UNIVERSIDAD AUTÓNOMA DE MADRID

Departamento de Biología Molecular

Facultad de Ciencias

DESARROLLO DE MÉTODOS COMPUTACIONALES BASADOS EN CO-EVOLUCIÓN PARA LA PREDICCIÓN DE INTERACCIONES ENTRE PROTEÍNAS

TESIS DOCTORAL

David Alejandro de Juan Sopeña

Madrid, 2015

UNIVERSIDAD AUTÓNOMA DE MADRID

Departamento de Biología Molecular

Facultad de Ciencias

DESARROLLO DE MÉTODOS COMPUTACIONALES BASADOS EN CO-EVOLUCIÓN PARA LA PREDICCIÓN DE INTERACCIONES ENTRE PROTEÍNAS

TESIS DOCTORAL

Memoria presentada para optar al grado de doctor en Ciencias por:

David Alejandro de Juan Sopeña

Centro Nacional de Investigaciones Oncológicas (CNIO)

Madrid, 2015

Dirigida por: Prof. Alfonso Valencia Herrera

cnio
Centro Nacional
de Investigaciones
Oncológicas

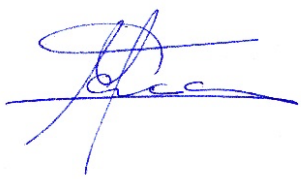
D. Alfonso Valencia Herrera, Director del Programa de Biología Estructural y Biocomputación,

CERTIFICA: Que D. David Alejandro de Juan Sopeña ha realizado el trabajo original de investigación "Desarrollo de métodos computacionales basados en co-evolución para la predicción de interacciones entre proteínas", bajo su dirección en el Centro de Nacional de Investigaciones Oncológicas, para la obtención de Grado de Doctor.

Que considera que dicho trabajo reúne las condiciones necesarias para su presentación y defensa entre tribunal en la Facultad de Ciencias de la Universidad Autónoma de Madrid.

Y para que así conste, a los efectos oportunos, firma el presente certificado.

Madrid, a 26 de Octubre de 2014



Prof. Alfonso Valencia Herrera

"Si supiera lo que hago, no lo llamaría investigación ¿no?"

Albert Einstein

AGRADECIMIENTOS

Ante todo quiero agradecer a mi director, Alfonso Valencia, la oportunidad que me brindó de entrar en su grupo y de compartir sus retos y conocimientos. Alfonso, muchas gracias por haberme dado el apoyo, los consejos, la confianza y el tiempo necesarios para realizar esta tesis. Me has enseñado el valor del trabajo, de la pasión por la ciencia y de la dedicación sin fin. Tú pusiste las primeras piedras sobre las que se construyó esta tesis y tus conocimientos y tu guía la han hecho posible. Te agradezco el haberme dado la oportunidad de pensar contigo y de aprender a disfrutar del reto de afrontar tantas preguntas nuevas con un fuerte espíritu crítico.

Después de Alfonso, mis mayores gracias son para Florencio Pazos, mi hermano mayor co-evolutivo. Gracias por tu paciencia, tu apoyo y tu tiempo cuando no lo merecía. Gracias por dejarme aprender sin límite de tus conocimientos, pero también de tu sentido común, tu humildad, tu escepticismo y tu claridad de ideas (aunque no se me haya pegado nada). Compartimos la pasión por una idea y esa idea me la metiste tú en la cabeza.

También quiero dar las gracias a Federico Morán por dejar que me "colara" en su laboratorio y mostrarme la existencia de este universo de la Biología Computacional.

Llegados a este punto, se me acumulan las personas a las que agradecer que hayan compartido conmigo la aventura de esta tesis. Sois tantos y tan importantes para esta tesis, que si no os organizo en grupos nadie llegará a leerse ni el resumen de esta memoria, pero como sabéis todos estáis en mayor o menor medida en cada uno ellos.

Pondré en primer lugar a mis amigos los discutidores: Antonio, Víctor, Dani, Fede, Simone, y Vera. Juntos compartimos "discusiones interminadas" sobre lo humano y lo divino (con alguno hasta más de lo segundo) y los mejores ratos de este negocio: la búsqueda de ideas. Ya sabéis, esos momentos delirantes en los que buscamos como posesos la moneda que algún mago tramposo nos escondió detrás de la oreja. Muchas gracias por compartir vuestros momentos geniales conmigo, por olvidaros de todo y entrar conmigo en ese fumadero de opio que es la pasión por entender.

En mi lista de agradecimientos con mayúsculas están mis amigos sabios: Michael, Ana, Alonso, David (Pisano), Enrique y Tirso. Me habéis regalado consejos y apoyo sin fin. Es un privilegio disfrutar de ambos y sin ellos nunca habría llegado hasta aquí. Habéis conseguido que respete vuestras opiniones (las científicas también) más que las mías mismas y por buenos motivos: son mejores. Gracias mil.

*Ahora les toca el turno a los que de verdad han dado el callo durante el tiempo de esta tesis: Osvaldo, Txema, David (Ochoa), Edu, Iakes, Beatriz, Ángel, Dorota, Raquel y últimamente sobre todo Juan. Ha sido un placer aprender juntos sobre co-evolución, pero también sobre todos los demás proyectos que hemos compartido. Hemos sudado sobre esas teclas que nos meten en códigos llenos de bichos con muy mala leche y hemos currado como c*br*n*s, que es lo que nos gusta. Gracias por luchar en las trincheras conmigo y ayudarme siempre que lo he necesitado (y es muy a menudo). Para mí esta parte siempre ha sido la más importante porque es de verdad.*

Para el resto de (ex)miembros del grupo (Gonzalo, José Manuel, José María, Paolo, Ángela, Florian, Luis, Fátima, Damien, Christian, Juan Antonio, Ramón, Almudena, Martin, Miguel, Manolo, Paulino, Jon, María, ...) muchas gracias por compartir conmigo tantas horas de descanso (y algunas de curro) y no pocas confidencias y consejos. Entre todos habéis hecho de este grupo mi casa. Con muchos de vosotros (y del resto) también comparto la pasión por el hockey, ese deporte loco del que no sabíamos nada hasta que Michael y Txema nos metieron la curiosidad (y algún que otro palo) en el cuerpo y que nos ayuda a desconectar cuando la cabeza no da para más (¡Hurra Pingüinos!).

Si no eres del grupo, tal vez deberías haber empezado por aquí.

Quiero dar mil gracias a mis pacientes amigos: Esteban, Rubén, Óscar, Ana, Milagros, Raquel, Julia y sobre todo a mi hermano. Gracias, por no haberme repudiado (motivos os he dado durante esta tesis) por rarito, plasta y desastre. Es genial ver que tenemos tantas cosas en común y poder compartir con vosotros la vida. Sin vuestro cariño, vuestra comprensión y vuestros pies en la tierra hace tiempo que llevaría la única camisa que me viene al cuerpo (la de fuerza).

Millones de gracias a mi familia y a mis padres por su amor, apoyo, paciencia y confianza desde que mi cabezón asomó al mundo. Soy quien soy por vosotros y si hay algo aquí de valor es tan vuestro como mío.

Sobre todo esta tesis es para mi madre y mi abuelo, dos personas increíbles que me enseñaron las cosas más importantes: a amar, a ser curioso, a reírme de mí mismo, a ser humilde y también cabezota. Me siento la persona más afortunada, por haber podido compartir con vosotros los mejores y los peores momentos. Sois todo lo bueno que encuentro en mí y espero no desmerecer todo el esfuerzo que pusisteis en hacer de este mendrugo un hombre de provecho.

Finalmente, Belén, esta tesis es más tuya que mía.

Tu energía sin fin consigue ponerme en marcha cada mañana.

Tu amor sin duda hace de cada día algo mejor que el anterior.

Tu paciencia sin límites me da la confianza que pide la vida.

Tu apoyo sin fisuras me ha traído hasta aquí.

¡Cómo no te voy a querer!

Os doy gracias a todos porque sois la suerte de esta tesis.

RESUMEN

La co-evolución es el proceso por el que las interacciones de agentes evolutivos (especies, proteínas, etc.) evolucionan acumulando cambios dirigidos por la selección natural en dichos agentes. Por tanto la co-evolución es un componente clave de la teoría de la evolución y es esencial para comprender las redes de interacciones de los agentes evolutivos. A menudo la co-evolución se manifiesta en la acumulación (casi) simultánea de cambios en los agentes que interaccionan. Esta dinámica evolutiva resulta en la evolución paralela en ambos agentes de los caracteres heredados responsables de la interacción y, ocasionalmente, en la de los agentes mismos. En consecuencia, el rastro de la co-evolución puede inferirse de las similitudes entre los árboles filogenéticos de los caracteres o agentes que interaccionan.

A nivel molecular, se han desarrollado métodos basados en co-evolución para predecir interacciones entre proteínas y contactos entre residuos de aminoácidos. En concreto, se han utilizado los parecidos entre árboles de proteínas para detectar interacciones entre ellas. Esta aproximación, denominada *MirrorTree* (*MT*), detecta una asociación significativa entre parecidos de árboles e interacciones de proteínas, pero también recupera similitudes altas para muchos pares de proteínas que no interaccionan.

El objetivo principal de esta tesis es desarrollar métodos computacionales basados en los parecidos entre árboles capaces de predecir interacciones entre proteínas con alta fiabilidad. Para ello se diseñaron estrategias para analizar el proteoma completo. En particular, se diseñaron análisis de correlaciones parciales para recuperar señales de parecidos evolutivos predictivos de interacciones funcionales. Así, cada similitud entre dos árboles se evaluó frente a los árboles del resto del proteoma. En esta aproximación, estos "otros árboles" se constituyen como variables externas portadoras de señales engañosas potencialmente responsables de la similitud entre árboles observada.

Es este marco se desarrollaron dos métodos diferentes: *ContextMirror* (*CM*) y *ContextMirror Global* (*CMG*). *CM* se diseñó para extraer similitudes compartidas por un pequeño número de proteínas que fueran potencialmente informativas de co-evolución de varias proteínas en grupo. La evaluación del rendimiento predictivo de *CM* en *Escherichia coli* muestra una clara mejoría comparada con *MT*. Más aún, el análisis de estos resultados demuestra un mayor potencial predictivo en grupos funcionales de proteínas, como complejos o rutas metabólicas.

En cambio, *CMG* se diseñó para extraer parecidos de árboles específicos de cada par de proteínas. Una evaluación en 23 especies bacterianas muestra que *CMG* supera claramente a *CM* y recupera predicciones más fiables. La comparación entre predicciones acertadas de *CM* y *CMG* muestra un solapamiento pequeño. Además, el solapamiento de predicciones acertadas de *CMG* en diferentes especies es también modesto. Sin embargo se observa un mayor solapamiento entre anotaciones funcionales más generales. En concreto la *fosforilación oxidativa*, el *transporte de membrana* y el *flagelo* están entre los procesos y estructuras más habitualmente señalados por las predicciones de *CMG* en diferentes especies.

En conjunto, tanto *CM* como *CMG* muestran una capacidad predictiva muy buena en especies bacterianas. Estos buenos resultados demuestran que el proteoma completo es un marco adecuado para analizar la co-evolución entre proteínas. Es más, el éxito de estas estrategias complementarias sugiere que la co-evolución ocurre a diferentes niveles de la organización funcional de las proteínas. Finalmente, *CM* y *CMG* son una combinación potente para la predicción de interacciones funcionales entre proteínas y la exploración de los procesos co-evolutivos en bacterias.

ABSTRACT

Evolution describes the natural selection driven accumulation of changes in evolutionary agents such as species and proteins. Co-evolution is the process by which interactions between these evolutionary agents evolve. Co-evolution is a key component of the theory of evolution and is essential for understanding the interaction networks of evolutionary agents. Often, co-evolution shows up as the (close to) simultaneous accumulation of changes in the interacting agents and results in the parallel evolution of inherited features responsible for the interaction in both agents. As a consequence, co-evolutionary traces can be inferred from similarities between the phylogenetic trees of the interacting features or agents.

At the molecular level, a wealth of co-evolution-based methods have been developed for predicting protein interactions and amino acid residue contacts. In particular, protein tree similarities have been used for detecting protein interactions. This approach, known as *MirrrorTree* (*MT*), detects a significant association between tree similarities and protein interaction, but it also retrieves high tree similarities for many non-interacting protein pairs.

The main goal of this thesis is the development of tree similarity based computational methods that are able to generate high quality protein interaction predictions. For this, strategies to take advantage of analysing the whole proteome were designed. More specifically, partial correlation analyses were designed to retrieve those tree similarity signals predictive of protein functional interactions. In fact, every tree-tree similarity was evaluated in the context of the protein trees for all proteins in a reference proteome. This approach regards 'other trees' as external variables carrying deceptive signals that are potentially responsible for the observed tree-tree similarity.

In this conceptual framework, two different methodologies were developed: *ContextMirror* (*CM*) and *ContextMirror Global* (*CMG*). *CM* was designed to extract tree similarities shared by a small number of proteins that were potentially informative of protein group co-evolution. Evaluation of the predictive performance of *CM* in *Escherichia coli* shows clear improvement when compared to *MT*. Moreover, analyses of these results show higher predictive power for functional groups of proteins, such as protein complexes or metabolic pathways.

In contrast, *CMG* was designed to extract tree similarities that were strictly specific to every protein pair. Evaluation of the performance of the *CMG* predictions in 23 bacterial species shows that *CMG* outperforms *CM*. Comparison of successful *CM* and *CMG* predictions in *E. coli* shows a small overlap between predictions. Similarly, successful predictions of *CMG* for different species also show modest overlaps. However, a higher overlap was observed at the level of more general functional annotations. Concretely, Oxidative Phosphorylation, transmembrane transport and flagellum are among the most commonly targeted processes and structures by *CMG* predictions in different species.

As a whole, both *CM* and *CMG* show very good predictive power for bacterial species. These results show that the whole proteome is a more suitable framework for analysing protein co-evolution. Moreover, the success of these complementary strategies suggests that co-evolution occurs at different levels of protein functional organization. Finally, *CM* and *CMG* are a powerful combination for the prediction of functional interactions and the exploration of co-evolutionary processes in bacterial species.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ABREVIATURAS

BBH: mejor resultado bidireccional (del inglés *Best Bi-directional Hit*).

CM: *ContextMirror.*

CMG: *ContextMirror Global.*

FDR: tasa de falsos descubrimientos (del inglés *Fold Discovery Rate*).

FN: falsos negativos.

FP: falsos positivos.

GO: ontología génica (del inglés *Gene Ontology*).

MT: *MirrorTree.*

MTe: *MirrorTree estandarizado.*

PC: *Perfiles Co-evolutivos.*

PCc: *Perfiles Co-evolutivos contraídos.*

ROC: característica operativa del receptor (del inglés *Receiver Operating Characteristic*).

SDP: posiciones determinantes de especificidad (del inglés *Specificity Determining Positions*).

TMT: *Tol-MirrorTree.*

TN: verdaderos negativo (del inglés *True Negatives*).

TP: verdaderos positivos (del inglés *True Positives*).

1. Introducción

1.1. Co-evolución: un concepto ecológico

Aunque esta tesis se engloba en el área de la co-evolución molecular, no se puede comprender sin la referencia constante a la co-evolución entre especies. Como ocurre en general con los conceptos evolutivos, la comprensión de la co-evolución comenzó en estudios ecológicos y aún tiene en ellos una de sus puntas de lanza más importantes. Así, comenzar con una mirada conceptual a lo que estamos aprendiendo de estos estudios de campo es la mejor manera de entender la dimensión de una idea tan fundamental como la de la co-evolución.

La idea de que las interacciones ecológicas entre especies pueden favorecer la aparición de dinámicas evolutivas de mutua adaptación, fue planteada por el propio Darwin en "El origen de especies" (Darwin, 1859) y posteriormente explorada en su siguiente libro "Sobre los varios diseños por los cuales las orquídeas británicas y extranjeras son fertilizadas por los insectos, y sobre los buenos efectos del entrecruzamiento" (Darwin, 1862).

Sin embargo, el término co-evolución no fue acuñado hasta 1964 por Ehrlich y Raven (Ehrlich & Raven, 1964) en su emblemático artículo sobre la interacción entre mariposas y plantas. En este artículo se define co-evolución como aquellas interacciones evolutivas halladas entre diferentes especies u organismos donde el intercambio de información genética entre las especies se asume mínimo o nulo. Este trabajo sirvió para incentivar el desarrollo de estudios y modelos de co-evolución entre especies, haciendo evidente su relevancia para la comprensión de los procesos evolutivos.

Entre las diferentes definiciones de co-evolución, cabe destacar la de John N. Thompson (Thompson, 1994). Según Thompson la co-evolución es el proceso de evolución recíproca entre especies que interaccionan dirigido por la selección natural. Esta definición implica que la conservación de la interacción entre dos especies a lo largo del tiempo establece el marco para que algunos rasgos de ambas especies se adapten mutuamente (co-adaptación). Es importante tener presente que en una interacción evolutiva el éxito reproductivo de las dos especies es mutuamente dependiente, por lo que cambios, en cualquiera de las especies, de los rasgos que determinan la interacción, afectarán a la supervivencia de ambas especies. Dicha dependencia no significa necesariamente que la interacción sea favorable para ambas especies aunque, lógicamente, deberá ser compatible con su supervivencia, o también la interacción desaparecerá.

Probablemente, el modelo más conocido e influyente de co-evolución sea la hipótesis de la Reina Roja propuesta por Van Valen en el contexto macroevolutivo de su Ley de las Extinciones (Van Valen, 1973). La formulación original de la hipótesis de la Reina Roja de Van Valen, es ilustrada por una escena de "Alicia a través del espejo" de Lewis Carrol, en la que la Reina Roja explica a Alicia que en su país hay que correr tanto como uno pueda para permanecer en el mismo sitio, ya que en él todo está en continuo movimiento. De forma similar, Van Valen propone que las especies están cambiando constantemente para poder mantenerse en un medio en el que todo cambia, incluyendo al resto de las especies con las que se relacionan. Esta hipótesis implica que la co-evolución entre especies que interaccionan debe ser la norma, aunque en la mayoría de los casos tal co-evolución sería difícil de apreciar, dado el elevado número de

interacciones que se dan simultáneamente. En estos casos se habla de co-evolución difusa (Fox, 1981), y suponen la generalización de la situación co-evolutiva arquetípica, en la que el carácter íntimo y continuado de las interacciones entre dos especies permite observar señales claras de co-evolución entre ellas. También se debe tener en cuenta que muchas de las interacciones ecológicas serán efímeras y por tanto no tendrán consecuencias detectables en la evolución de las correspondientes especies. Así, estas interacciones ecológicas lábiles no darán lugar a interacciones evolutivas, ni por tanto a co-evolución.

1.1.1. Co-evolución: cambiar para conservar

Uno de los méritos de la hipótesis de la Reina Roja de van Valen es la forma intuitiva en que conecta dos componentes esenciales y aparentemente contradictorios de la co-evolución: la conservación y el cambio. De hecho, la co-evolución se puede entender como la combinación de ambos fenómenos operando a diferentes niveles.

Como se ha comentado, los procesos de co-evolución entre pares de especies implican la conservación de una interacción importante para al menos una de las especies que interaccionan. La conservación es un concepto de larga tradición en los estudios evolutivos, pues es una señal inequívoca de la importancia del carácter conservado. Sin embargo, dicha conservación se suele entender (especialmente a nivel molecular) como la ausencia de cambio. De hecho, esta resistencia a cambiar es la señal más evidente de una fuerte selección negativa en la que aquellos individuos que introducen cambios en el carácter se ven penalizado, y sus alternativas, menos exitosas, tenderán a ser progresivamente eliminadas de la población.

En el contexto de la co-evolución, esta selección negativa se ve reflejada en la conservación de la interacción. La ruptura de dicha interacción ha de ser perjudicial para al menos una de las especies y es la presión de selección sobre las especies beneficiarias la que se impone al efecto perjudicial que pueda tener sobre otras especies en la interacción. En cierto sentido se puede hablar de presión de selección sobre el ensamblado de especies que co-evolucionan (es decir sobre un ecosistema evolutivo). De hecho, la conservación de una interacción, aunque no beneficie a todas las especies implicadas, demuestra su éxito como forma de interacción evolutiva.

El escenario más ilustrativo de esta situación es el de las interacciones antagonistas. En él se suelen dar carreras armamentísticas, donde sólo constantes cambios en el depredador pueden contrarrestar los cambios debilitantes de la interacción en la presa, permitiendo mantener la interacción (Cott, 1940). Sin embargo, hasta en las interacciones mutualistas, las especies siguen acumulando cambios como parte de su mutua adaptación y de la adaptación de su interacción a otros factores (como otras especies que puedan parasitar a una de ellas aprovechando la interacción (Currie *et al.*, 1999)).

De hecho, en todo ecosistema se da una constante evaluación de soluciones alternativas en el contexto de unas condiciones también cambiantes. Es en este punto en el que la hipótesis de la Reina Roja es muy útil para comprender que incluso la conservación de una interacción evolutiva requiere de su adaptación a los cambios introducidos en las especies que interactúan, así como en otros elementos con los que éstas interaccionan, sea cual sea su naturaleza.

Dado que la co-evolución implica un cambio que posibilita la conservación de las interacciones evolutivas, es interesante entender su posible influencia en los procesos de divergencia y especiación. Son particularmente sugerentes los trabajos recientes que

aluden a la capacidad de los procesos de co-evolución para promover la aparición de innovaciones evolutivas (Marston *et al.*, 2012; Meyer *et al.*, 2012; Zaman *et al.*, 2014), así como aquellos que plantean su posible papel como motor de procesos de especiación en interacciones antagonistas o competitivas (Thompson, 1994; Hembry *et al.*, 2014).

1.1.2. Detectando la co-evolución entre especies

Un aspecto clave para establecer la existencia de procesos co-evolutivos entre especies es definir qué señales se consideran suficientes para determinar su presencia. Existe una larga tradición (desde el mismo Darwin) de utilizar cambios de caracteres coordinados entre las especies que interaccionan para investigar la evolución de dicha interacción. Un ejemplo clásico es la relación entre las longitudes de nectarinos y proboscis de diferentes parejas de orquídeas y sus polinizadores, principalmente mariposas, (Darwin, 1862). La correspondencia de estas longitudes define la viabilidad y exclusividad de las correspondientes interacciones ecológicas, ya que establece si el polinizador podrá alcanzar el néctar que usa como alimento y si para ello deberá recoger el polen cuyo transporte es esencial para la fecundación de las orquídeas. La relevancia de dichos caracteres para la interacción queda reflejada por su propia co-variación a lo largo de muchas parejas de especies. En último término, esta co-variación muestra la fuerte co-dependencia de dichos caracteres y de las especies que interaccionan, sugiriendo una historia evolutiva paralela de dichos caracteres.

Más adelante, con el desarrollo de la filogenia molecular, se descubrió que, en algunos casos, dicha co-variación podía observarse de forma más global en la evolución de las especies que interaccionan muy intensamente. Por ejemplo, se observaron paralelismos en las tasas de divergencia e incluso procesos de especiación coetáneos en casos de mutualismos y parasitismos obligados, en los que la interacción entre las especies es necesaria para la supervivencia de al menos una de ellas (Kellogg, 1896; Fahrenholz, 1913). Esto derivó en la llamada Regla de Fahrenholz (Eichler, 1948), según la cual la filogenia de los parásitos eran un reflejo de la filogenia de su hospedador, para la se han documentado numerosos casos (Hafner & Nadler, 1988). No obstante, la similitud en las historias evolutivas de las especies que interaccionan dista mucho de ser la situación general (Thompson, 2013). Además, estos signos de interacción, al afectar a aspectos muy generales de la evolución de las especies, no necesariamente se asocian exclusivamente a procesos de co-adaptación (Althoff *et al.*, 2013). Por ejemplo, la co-especiación se suele vincular a interacciones con transmisión vertical de la interacción de padres a hijos, con poca transmisión horizontal entre especies o incluso miembros de una comunidad.

Por otro lado, los paralelismos filogenéticos, aunque pueden estar en parte relacionados con procesos de co-adaptación, es muy posible que también reflejen la medida en que cambios externos a la interacción han influido de forma semejante a las especies que interaccionan (Althoff *et al.*, 2013). Por ejemplo, es evidente que los cambios en los tamaños de las poblaciones de una de las especies, a menudo, llevará asociados cambios en la otra especie dependiente, o beneficiaria de la interacción (por ejemplo en situaciones de hambruna). Por tanto, en el caso de interacciones muy estables, las especies que co-evolucionan pueden conformar (o acercarse a ser) una unidad de selección. Esta situación, aunque es igualmente consecuencia de la interacción, no puede considerarse señal de la co-adaptación entre especies, sino de la evolución de la interacción, que resulta en historias evolutivas semejantes en las especies que co-evolucionan. El estudio de estas dinámicas poblacionales se ha beneficiado del desarrollo de la teoría coalescente (Rosenberg & Nordborg, 2002) que establece un

marco estocástico para la inferencia retrospectiva de los procesos por los que ha pasado una población a partir de la diversidad genética en la población actual. Este modelo supone una aplicación extremadamente potente del conocimiento acumulado sobre la genética de poblaciones. Es el criterio seguido en esta memoria, que estos casos de adaptación conjunta también son dinámicas co-evolutivas, las cuales incluyen pero no se limitan a fenómenos de co-adaptación (Juan *et al.*, 2008a) (ver Discusión).

Otro factor a tener en cuenta por su efecto distorsionador de las filogenias moleculares es el reparto incompleto de linajes (Degnan & Rosenberg, 2009). Esta situación, también desvelada mediante la teoría de coalescencia, implica una segregación imperfecta de los diferentes alelos de un gen en los diferentes linajes evolutivos. Por ejemplo, cuando se dan dos (o más) especiaciones muy seguidas temporalmente, una retención dispar de los alelos ancestrales en los linajes resultantes, puede producir que especies próximas presenten alelos diferentes, mientras que especies lejanas mantienen el mismo alelo. En estos casos, la filogenia reconstruida a partir de este gen muestra una mayor similitud entre las especies que retienen el mismo alelo a pesar de ser evolutivamente más distantes. Efectos de este tipo pueden afectar a la filogenia de multitud de genes, dando lugar a árboles de especies erróneos (Degnan & Rosenberg, 2006) y afectar negativamente a la detección de señales co-evolutivas.

1.1.3. Escenarios co-evolutivos

Durante las últimas décadas se han multiplicado el número de trabajos que han establecido la relevancia de diversos escenarios de co-evolución entre especies. Estos escenarios incluyen las interacciones depredador-presa y las patógeno-hospedador como ejemplos de interacciones antagonistas, en las que, como se ha comentado, se dan auténticas carreras armamentísticas de competencia constante entre medidas ofensivas y defensivas. La presión de selección asociada a estas carreras armamentísticas parece acelerar la incorporación de innovación evolutiva y puede haber sido la fuente de gran parte de la diversidad fenotípica de las especies existentes (Paterson *et al.*, 2010).

Así mismo, también se observa co-evolución en interacciones mutualistas, en las que las especies que interactúan se benefician de la interacción. En mutualismos la evidencia de la relación entre co-evolución y aceleración evolutiva parece contradictoria, probablemente reflejando una situación en que las interacciones son estables a cambios moderados, pero pueden ser rápidamente desplazadas a otras formas de interacción por cambios más drásticos en las condiciones (Hembry *et al.*, 2014). Es particularmente interesante que estas interacciones específicas parecen tener la capacidad de atraer a nuevas especies que convergen hacia las características de uno de los miembros de la interacción, estableciendo una relación de competencia por la interacción o de parasitismo de la misma (Thompson, 2013). Este escenario parece estar en la base del establecimiento de redes ecológicas y evolutivas más complejas y puede haber actuado como semilla de muchas de las redes ecológicas conocidas.

Hasta ahora se han discutido los casos en que dos o más especies interaccionan de forma permanente a lo largo del tiempo, caso ideal para la detección de señales claras de co-evolución. Sin embargo, como se ha explicado previamente, gran parte de la co-evolución puede ser difusa debido al número elevado y cambiante de interacciones a lo largo de diferentes ecosistemas y momentos. De hecho, es probable que en la mayoría de los casos dichas interacciones sean relativamente lábiles, dependiendo de las fluctuaciones del tamaño de las poblaciones, así como de la interferencia de factores externos (*e.g.* factores ambientales) (Thompson, 2013).

La exploración de los diferentes escenarios de co-evolución entre especies ha acompañado a la creciente consciencia del elevado grado de interdependencia existente entre las especies que comparten un ecosistema dado. De hecho, la complejidad y densidad de estas redes de interdependencias han llevado a considerarlas como sistemas complejos, dando lugar al establecimiento de una nueva disciplina: la ecología de sistemas (Tansley, 1935).

1.1.4. Ecología de sistemas y la teoría del mosaico co-evolutivo

La ecología de sistemas (como otras disciplinas apellidadas "de sistemas") surge de la aplicación de los conceptos y métodos del campo de los sistemas complejos a otro ámbito de estudio (en este caso, las interacciones entre especies y con su medio). La definición de qué es un sistema complejo es controvertida, pero hay algunas propiedades que parecen ser características de ellos, tales como el estar compuestos de un elevado número de componentes conectados en una intricada red de interacciones, su robustez a cambios aleatorios externos o internos (Albert *et al.*, 2000; Montoya *et al.*, 2006), o su capacidad para ser fuente de propiedades emergentes (Foote, 2007). Estas propiedades emergentes se han definido como aquellas que no pueden explicarse únicamente a partir de las propiedades de los elementos que lo componen, sino que se derivan de la propia estructura y dinámica de las relaciones entre elementos y sus consecuencias en el comportamiento del sistema como un todo (Foote, 2007).

Establecer estas propiedades emergentes es a menudo difícil. Aunque es sencillo definir como tal cualquier propiedad establecida a nivel de sistema, resulta complicado determinar la medida en que ésta es o no 'explicable' desde los elementos del sistema, así como encontrar ejemplos fácilmente medibles, cuya dinámica se pueda analizar. Entre las propiedades propuestas como emergentes en la ecología de sistemas se encuentran la aparición de patrones auto-organizativos geográficos, temporales o estructurales (Parrish & Edelstein-Keshet, 1999). Desde nuestro punto de vista, es interesante que los propios procesos co-evolutivos parecen ser capaces de actuar como promotores de complejidad (Thrall *et al.*, 2007), afectando a la estructura y dinámica de los ecosistemas (Bascompte *et al.*, 2003; Bastolla *et al.*, 2009; Thébault & Fontaine, 2010). Desde esta perspectiva, cabe plantearse si los procesos de especiación, así como la propia divergencia adaptativa no son de hecho propiedades emergentes de los sistemas ecológicos. En esta línea, no es casual que una de las teorías evolutivas más exitosas de las últimas décadas haya sido la teoría del mosaico geográfico co-evolutivo (Thompson, 2005).

La teoría del mosaico geográfico co-evolutivo, propone que la desigual distribución geográfica de distintos ambientes, interacciones ecológicas y poblaciones de la especies que interaccionan, conforma un mosaico geográfico de selección. Así mismo, establece que estas estructuras geográficas de la distribución de poblaciones de diferentes especies es el escenario clave para entender las dinámicas de los procesos co-evolutivos. Este marco conceptual acerca la biología evolutiva y la ecología a partir de reconocer el importante papel que juega la riqueza de escenarios diferentes pero geográficamente conectados que se están desarrollando en un momento dado. De esta manera la teoría del mosaico geográfico co-evolutivo ha contribuido a establecer un marco de trabajo mucho más realista para el estudio de las interacciones ecológicas que conllevan interacciones co-evolutivas y de su papel en el establecimiento de la diversidad de ecosistemas existentes (Thompson, 2013).

1.1.5. Una jerarquía de niveles evolutivos

En el contexto de la co-evolución entre especies se han acumulado muchos casos en la literatura de co-evolución gen-a-gen (Brown & Tellier, 2011). En estos casos la interacción entre dos especies que co-evolucionan se concentra en dos genes (uno por especie) que son particularmente relevantes para la interacción entre dichas especies. Se podría decir que al menos una de las funciones de dichos genes es mantener la interacción entre ambas especies. De hecho, suele ser posible detectar señales de co-adaptación molecular entre dichos genes. No obstante, parece evidente que en la mayoría de los casos la co-evolución debe darse entre grupos de genes de ambas especies. Estas interacciones moleculares entre especies nos recuerdan que la co-evolución a nivel de especies se fundamenta a su vez en la co-evolución a nivel molecular.

La co-evolución gen-a-gen sugiere la conveniencia de considerar una jerarquía anidada de niveles evolutivos que nos ayude a comprender y analizar diferentes escenarios co-evolutivos. Para los intereses de esta tesis, se puede considerar una jerarquía anidada simplificada (ver Figura 1). Dicha jerarquía considera como nivel superior el nivel de especies discutido hasta el momento. Dentro de él, de forma similar a un juego de muñecas rusas, se hallarían el nivel de genes/proteínas y dentro de éste el de residuos de aminoácidos en posiciones específicas de una proteína.

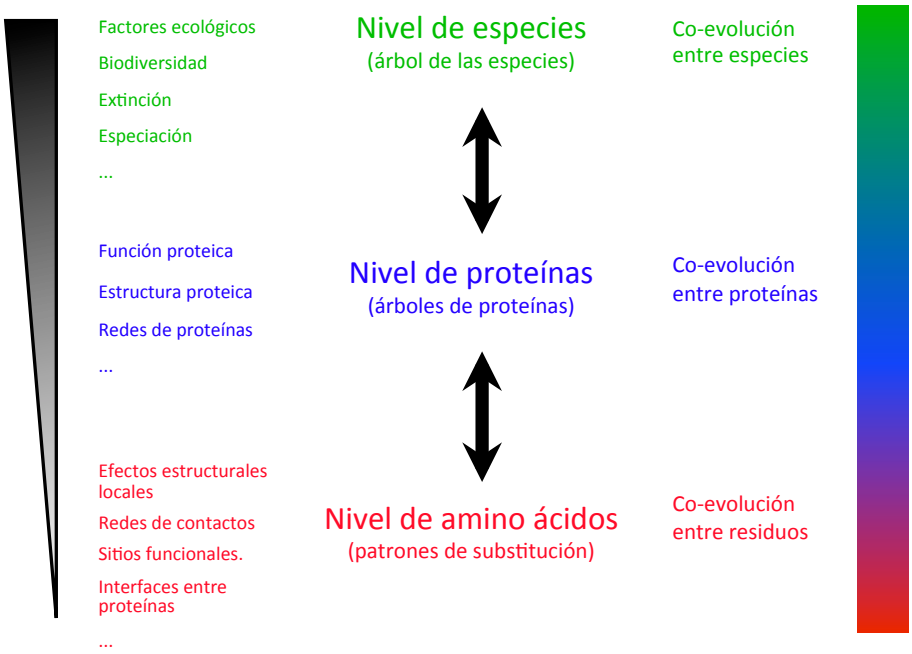


Figura 1. Jerarquía simplificada de niveles evolutivos.

Una jerarquía más completa podría incluir otros niveles como los de ecosistemas, poblaciones de individuos, complejos macromoleculares, los diversos elementos codificados en el genoma que no son proteínas, así como transcritos, codones, nucleótidos, etc. Igualmente debería distinguir entre gen y proteína, por no haber una equivalencia completa en la información que representan*.

La jerarquía simplificada de niveles evolutivos expuesta, se enfoca en tres capas diferentes pero inter-relacionadas en las que pueden darse procesos co-evolutivos.

* Sin embargo, en esta memoria se aludirá a la evolución de genes y proteínas en virtud de la información

Como se ha explicado, la observación directa de los ecosistemas ha permitido el estudio de dichos procesos al nivel de especies desde el mismo establecimiento de la teoría de la evolución. Sin embargo, los niveles microscópicos de proteínas y sus residuos, no fueron accesibles al estudio directo hasta la explosión más reciente de las técnicas en biología molecular (Morange & Cobb, 2000). En particular, las técnicas de secuenciación masiva han proporcionado ingentes cantidades de información referente a las dotaciones génicas de una gran variedad de especies (unas 7.500 especies completamente secuenciadas y publicadas en Octubre de 2015, según *GOLD* (Reddy *et al.*, 2015)). Como se verá a continuación, esta información es extremadamente valiosa para el estudio de la co-evolución (y de la evolución en general) a nivel molecular.

Así mismo en el nivel de especies, si bien se han acumulado los casos de pares de especies que co-evolucionan, también se ha hecho cada vez más evidente la existencia de niveles de mayor complejidad ecológica. Aunque alcanzar un mejor entendimiento de las dinámicas de los sistemas ecológicos es ahora más acuciante que nunca, el estudio evolutivo de estos sistemas que involucran a un gran número especies y localizaciones es logísticamente muy complejo (Thompson, 2013). La exploración de diferentes niveles moleculares de complejidad es igualmente esencial, aunque mucho más accesible. Esto hace de la co-evolución molecular un campo muy interesante para mejorar nuestra comprensión de la co-evolución en sistemas complejos (Thomson, 2013; Carmona *et al*., 2015). En este contexto se explicará cómo los métodos desarrollados en el estudio de la co-evolución a nivel molecular han avanzado en el tratamiento de las redes de elementos en co-evolución, así como en la detección de grupos de elementos que evolucionan de forma coordinada.

1.2. Co-evolución a nivel de residuos

La disponibilidad de las primeras secuencias de genes y proteínas proporcionó la oportunidad de realizar las también primeras reconstrucciones filogenéticas de la evolución molecular de algunos genes (Jukes, 1963; Pauling & Zuckerkandl, 1963). Estos estudios pioneros estuvieron dedicados fundamentalmente a comprender la evolución de las especies y a añadir evidencias moleculares al estudio de los registros fósiles que habían dejado muchas preguntas por responder. Esta etapa apasionante del estudio de la evolución de las especies llega hasta nuestros días, contribuyendo también a nuestra comprensión de la co-evolución entre especies (Thompson, 2013).

Sin embargo el estudio de la evolución molecular ha resultado ser mucho más complejo e interesante de lo esperado, ya que no sólo proporciona información sobre la evolución de las especies, sino también de la evolución de cada una de los genes como entidades evolutivas por derecho propio. Esta "personalidad evolutiva propia" de los genes añade importantes dificultades a su uso en el estudio de la filogenia de las especies, lo que ha derivado en el desarrollo de métodos que emplean la combinación de muchos genes diferentes para compensar esta situación (Degnan & Rosenberg, 2009).

Entre las observaciones que trajeron aquellos primeros análisis se encuentra la constatación de que las diferentes posiciones de la secuencia de una proteína también evolucionan de forma diferente. La construcción (entonces manual) de los primeros alineamientos de secuencias homólogas permitió observar la mayor conservación de determinadas posiciones asociadas con papeles importantes en la estructura y/o función de dichas proteínas (una fuente sin precio de información evolutiva) (Browne *et al.*, 1969; Schlesinger & Goldstein, 1975). Así mismo trabajos pioneros observaron que posiciones que cambiaban de forma coordinada en conjuntos de secuencias alineadas estaban próximas espacialmente en algunas de las primeras estructuras de proteínas

resueltas (Altschuh *et al.*, 1987; Altschuh *et al.*, 1988). Estas posiciones recibieron el nombre de mutaciones correlacionadas o coordinadas. El desarrollo de los primeros métodos capaces detectar dichas posiciones automáticamente (Korber *et al.*, 1993; Göbel *et al.*, 1994; Neher, 1994; Shindyalov *et al.*, 1994; Taylor & Hatrick, 1994), permitió confirmar estas observaciones en un número mayor de familias de proteínas (Göbel *et al.*, 1994; Taylor & Hatrick, 1994). Estos trabajos generaron grandes expectativas para la predicción de estructuras tridimensionales usando contactos predichos a partir de información de secuencia, ya que la secuenciación era (y sigue siendo) mucho más sencilla y barata que la resolución de estructuras tridimensionales.

1.2.1. Detección de cambios correlacionados en pares de residuos de proteínas

Las metodologías desarrolladas para la detección de mutaciones correlacionadas a partir de alineamientos múltiples de secuencias se pueden clasificar en dos clases principales: aquellas que se basan en el uso de la información mutua (Korber *et al.*, 1993; Tillier & Lui, 2003; Martin *et al.*, 2005; Dunn *et al.*, 2008) y las que emplean la correlación de los cambios observados (Göbel *et al.*, 1994; Neher, 1994; Taylor & Hatrick, 1994; Fares & McNally, 2006) (ver Figura 2). Ambas aproximaciones, buscan detectar señales de paralelismos en los perfiles mutacionales de cada par de posiciones para inferir la presencia de co-evolución y por tanto de contactos físicos. Estas metodologías, aunque demostraron la relación entre co-evolución e interacción, sólo obtuvieron un éxito moderado en la predicción de contactos.

Al abrigo de la introducción de las técnicas de aprendizaje automático en la biología computacional en la década de los noventa, se desarrollaron un gran número de métodos computacionales dedicados a la predicción de diferentes características proteicas (Rost & Sander, 1993; Rost *et al.*, 1995; Baldi *et al.*, 1999; Li *et al.*, 1999). Entre estas características se encontraba la predicción de contactos entre residuos de aminoácidos de la misma proteína. Una de las fuentes de información más útiles en esta tarea fue la información evolutiva extraída de alineamientos de secuencias. Ésta incluía tanto la conservación de posiciones, como la correlación entre perfiles mutacionales de pares de posiciones a lo largo de la familia de proteínas de la secuencia problema (Fariselli *et al.*, 2001; Punta & Rost, 2005). Una aproximación que más tarde también se empleó para la predicción de estructuras mediante la detección de homólogos remotos (Jones, 1999; Fischer, 2000; Juan *et al.*, 2003) (ver Figura 2). Sin embargo, estas aproximaciones alcanzaron su techo en la predicción de contactos sin llegar a ser capaces de producir modelos muy fiables y de alta resolución.

A lo largo de los años, se han identificado varios problemas para la detección de mutaciones correlacionadas predictivas de contactos. Entre ellos, cabe destacar la existencia de un sesgo producido por el hecho de que las secuencias del alineamiento, al ser homólogas, no son independientes entre sí (Fodor & Aldrich, 2004). Esto implica que parte de la señal detectada proviene de la estructura de similitudes entre secuencias, reflejo a la diferente divergencia evolutiva entre ellas. La identificación y eliminación de esta señal evolutiva han sido el objetivo de una serie de soluciones metodológicas, que se han ido incorporando y mejorando posteriormente (Tillier & Lui, 2003; Martin *et al.*, 2005; Dunn *et al.*, 2008). Éste y otros problemas identificados y afrontados por la comunidad han contribuido a una mejora paulatina de la capacidad de detectar contactos a partir de la señal co-evolutiva.

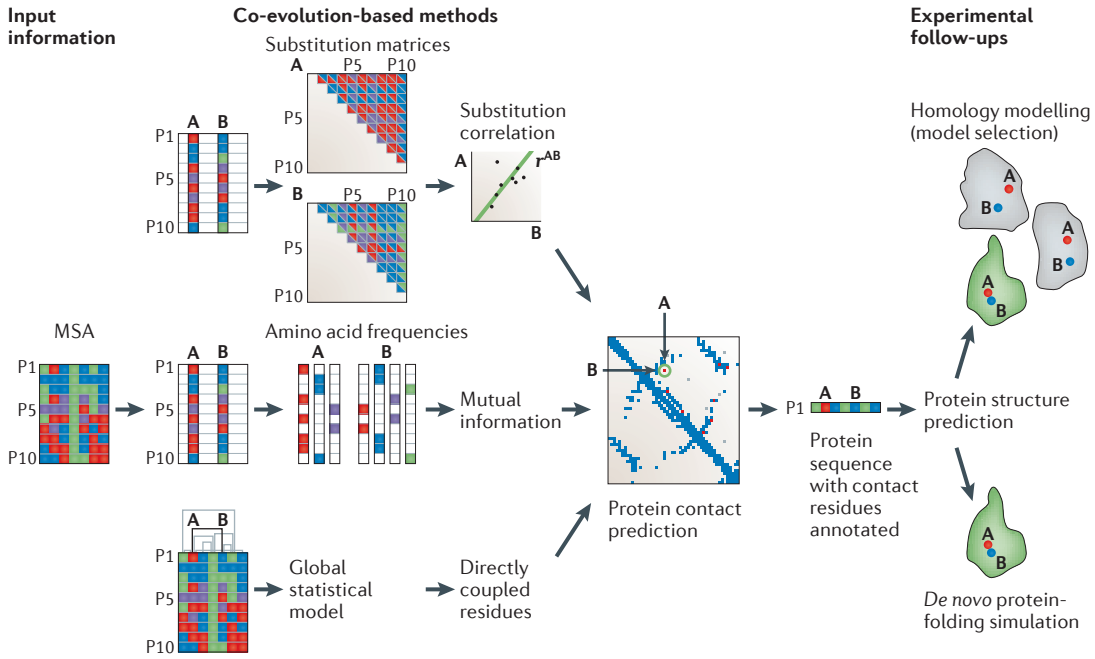


Figura 2. Aplicación de la co-evolución entre pares de residuos a la predicción de estructuras de proteínas (adaptada de (Juan *et al.*, 2013)). Se representan tres aproximaciones diferentes para la detección de co-evolución entre residuos que producen predicciones de contactos que pueden ser utilizadas para seleccionar modelos por homología o para la predicción de estructura *de novo*.

A pesar de las mejoras obtenidas en éste y otros ámbitos de la predicción de contactos, la calidad de las predicciones hasta hace poco seguía sin ser lo suficientemente buena para cumplir la expectativa de asistir con robustez a la predicción *de novo* de estructuras de proteínas. Sólo recientemente se ha podido alcanzar este hito gracias a la mejora en los modelos estadísticos utilizados para desentrañar la señal co-evolutiva (para una revisión detallada del campo ver (Juan *et al.*, 2013)).

1.2.2. Desentrañando el acoplamiento directo en la evolución de pares de residuos

Ya en los noventa Lapedes *et al.* (Lapedes *et al.*, 1999) señalaron un problema esencial de las redes de mutaciones correlacionadas que la Física había abordado previamente en otro contexto (Mézard *et al.*, 1986). Las redes de elementos interdependientes presentan similitudes en el comportamiento de los elementos que no se deben a la interdependencia existente entre ellos, sino que son un reflejo de interdependencias con otros elementos cuya similitud se transmite a lo largo de la red. La acumulación de similitudes "indirectas" es un factor de confusión importante a la hora de reconstruir la red de interdependencias "reales" a partir de las similitudes en el comportamiento de sus elementos. En este trabajo los autores también propusieron una aproximación basada en la idea de recuperar la distribución de inter-dependencias más sencilla (con máxima entropía) capaz de explicar la red de similitudes observada (Jaynes, 1957b; a). Sin embargo, dada la complejidad computacional utilizaron un "ejemplo de juguete", no llegando a demostrar su utilidad en casos reales.

En un trabajo posterior, en 2002, Lapedes *et al.* aplicaron su aproximación a algunas proteínas reales con buenos resultados (Lapedes *et al.*, 2002) resolviendo los problemas de complejidad computacional mediante el uso de heurísticas aplicadas a problemas

semejantes en otros sistemas (Berger *et al.*, 1996). Lamentablemente no llegaron a publicar sus resultados en ninguna revista haciendo su manuscrito públicamente disponible a través de un enlace permanente de internet. Como consecuencia estos resultados pasaron desapercibidos para la comunidad. En 2011 Weigt *et al*. desarrollaron de forma independiente una metodología similar y demostraron el potencial predictivo de la co-evolución (Weigt *et al.*, 2009), lo que ha derivado en una serie de desarrollos posteriores que han hecho progresar considerablemente este campo (Marks *et al.*, 2011; Morcos *et al.*, 2011).

Estas metodologías pretenden extraer las co-dependencias directas que explican la matriz de co-dependencias aparentes observadas entre las posiciones de un alineamiento múltiple (calculadas a partir de las frecuencias de co-aparición de residuos). En este sentido, todos estos métodos están basados en obtener un equivalente para variables categóricas de la matriz de correlaciones parciales para las continuas, capaz de explicar la distribución de valores de partida. La solución directa a este problema supone invertir la matriz de covarianzas (Whittaker, 2009).

Sin embargo, la matriz de partida, para los alineamientos múltiples, presenta una serie de problemas que dificultan su inversión, tales como demasiados parámetros para recuperar una solución única (surgidos de la expansión de las posiciones a tipo de residuos por posición), o el insuficiente muestreo (falta de secuencias) (revisados en (Stein *et al.*, 2015)). En esta situación, invertir la matriz de co-varianza de partida, no es posible. Estos problemas se pueden evitar introduciendo una técnica de fijado de calibre para reducir el número de parámetros y una corrección en la matriz de partida que simula un muestreo mayor de secuencias. A continuación se procede a inferir los parámetros del modelo de interacciones (informativos de las interacciones directas) que ha podido dar lugar a la matriz de frecuencias corregida (problema inverso). Este modelo se recupera mediante un ajuste de sus parámetros que busca la explicación más sencilla (máxima entropía) a las frecuencias observadas de residuos en las posiciones del alineamiento (Lapedes *et al.*, 1999; Weigt *et al.*, 2009; Morcos *et al.*, 2011; Jones *et al.*, 2012; Cocco *et al.*, 2013; Ekeberg *et al.*, 2013). Finalmente, la matriz de estos parámetros que corresponde a la inferencia de la inversión de la matriz de partida, se utiliza para establecer medidas de la co-dependencia directa entre los residuos (semejantes a la correlación parcial para variables continuas). Estas medidas, pueden incluir correcciones por el sesgo de fondo asociado a la no independencia de las secuencias discutido anteriormente (Jones *et al.*, 2012; Ekeberg *et al.*, 2013) (para una revisión de este marco conceptual ver (Stein *et al.*, 2015)).

La aplicación de estas metodologías a la predicción de contactos ha resultado en una reducción espectacular en el número de predicciones incorrectas, permitiendo obtener un número mucho menor, pero mucho más fiable de predicciones (Lapedes *et al.*, 2002; Weigt *et al.*, 2009; Morcos *et al.*, 2011).

1.2.3. Predicción de estructura de proteínas asistida por co-evolución

La mejoría en la calidad de las predicciones de contactos basadas en co-evolución ha permitido incorporarlas como restricciones tridimensionales en la predicción *de novo* (a partir de la secuencia) de estructura de proteínas. La incorporación de estas restricciones ayuda a acotar el espacio de soluciones a explorar por simulaciones de dinámica molecular, lo que resulta en muy buenos modelos estructurales (Marks *et al.*, 2011; Hopf *et al.*, 2012; Sułkowska *et al.*, 2012).

Actualmente la predicción *de novo* de estructuras de proteínas asistida por co-evolución se ha convertido en una alternativa interesante al diseño por homología cuando éste no es aplicable. En el diseño por homología, se aprovecha la existencia de la estructura resuelta de una proteína (o fragmento) homóloga de la proteína problema para reducir el espacio de posibles estructuras a aquellas similares a las homólogas. Para ello, se usan estas estructuras como moldes en los que acomodar (con pequeñas modificaciones) la proteína problema. Esta técnica proporciona predicciones estructurales muy precisas cuando se pueden detectar homólogos con estructura conocida.

Un punto importante es que los nuevos métodos para la detección de co-evolución entre residuos requieren un gran número de secuencias no redundantes (entre 500 y 1000 secuencias con una redundancia menor del 80% de identidad) (Morcos *et al.*, 2011). Este escenario delimita la capacidad de predicción de contactos con alta fiabilidad a familias de proteínas relativamente grandes y antiguas. Por lo tanto, al menos con estos métodos, parece improbable la detección precisa de señales recientes o que sólo afecten a una pequeña proporción de secuencias.

Como consecuencia, el ámbito de aplicación de la co-evolución entre residuos a la predicción estructural se concentra en casos donde existen muchas secuencias homólogas, pero ninguna de ellas está resuelta estructuralmente. Por ejemplo, ésta es la situación de las proteínas transmembrana, donde la dificultad para obtener cristales de buena calidad mantiene bajo el número de estructuras resueltas y donde existen familias enormes cuya expansión ha sido importante en la evolución de organismos pluricelulares (Hopf *et al.*, 2012; Hayat *et al.*, 2015). Así la predicción de interacción asistida por co-evolución se está aplicando con muy buenos resultados a una variedad de casos donde el diseño por homología aún no es aplicable (Marks *et al.*, 2011; Hopf *et al.*, 2012; Sułkowska *et al.*, 2012; Hayat *et al.*, 2015; Hopf *et al.*, 2015).

1.2.4. Predicción de modos de interacción entre proteínas

Aunque hasta ahora se ha tratado la co-evolución entre pares de residuos de la misma proteína, las metodologías comentadas son igualmente aplicables para la detección de co-evolución entre residuos de proteínas diferentes que interaccionen. Esta aplicación merece una mención aparte, tanto por su relevancia como por algunos aspectos característicos de sus análisis.

La predicción de interfaces de interacción entre proteínas es una de las aplicaciones más prometedoras de los métodos de co-evolución. En realidad, las proteínas raramente actúan de forma independiente, sino que deben coordinarse con otras moléculas para realizar sus funciones en el momento, lugar y modo adecuados para el organismo. De esta manera, se forman complicados entramados de redes de regulación y máquinas macromoleculares que presentan importantes co-dependencias moleculares. Comprender mejor la forma en que estos entramados se reflejan en las historias evolutivas de las proteínas nos ofrece la posibilidad de detectar interacciones cuya mutua adaptación ha sido clave para el funcionamiento de los organismos a lo largo de la evolución. De hecho esta información debería permitir predecir y entender mejor la interfaz, la orientación (modo) y la dinámica de dichas interacciones. Este conocimiento vendrá a completar la información experimental y computacional acumulada para diferentes interactomas (conjunto de interacciones de proteínas que se dan en los individuos de una especie), y podría contribuir a mejorar nuestra capacidad de intervenir en dichas interacciones con fines terapéuticos o biotecnológicos.

Desde sus inicios la predicción de interfaces se ha visto limitada esencialmente por dos factores: la falta de conservación de las regiones y la falta de especificidad de la información disponible. La falta de conservación implica que, a diferencia de lo que ocurre con el núcleo de la proteína o con los sitios de unión a ligandos pequeños, los residuos de las interfaces entre proteínas está relativamente poco conservados entre homólogos (aunque se conserve la interacción) (Caffrey *et al.*, 2004). Esta situación puede deberse a varias causas como: la conservación de propiedades más globales que los residuos individuales en forma de co-evolución de pares o grupos de residuos; la responsabilidad de distintas combinaciones de residuos en la afinidad o estabilidad de la interacción lo largo de la evolución; o la necesidad de incorporar cambios que se adapten a las condiciones de cada especie.

Por otro lado, las características de los residuos en la interfaz (incluyendo la conservación evolutiva o la hidrofobicidad) carecen de información acerca de cuáles son los pares o grupos de residuos que interaccionan entre ambas proteínas. Esta información, en el mejor de los casos, puede apuntar a la relevancia de una cierta región de la proteína que podría estar asociada a la presencia de una interfaz de interacción. Sin embargo, no permite discriminar entre distintas orientaciones de la interacción y ni siquiera determinar en cuál de las diferentes interacciones de la proteína participa.

En este contexto es particularmente útil establecer el grado de co-dependencia entre pares de residuos concretos de proteínas determinadas. Esto hace de la co-evolución una herramienta valiosa para superar las limitaciones asociadas al uso de información sobre los residuos considerados individualmente. Así detectar una señal co-evolutiva clara entre posiciones de dos proteínas diferentes, nos podría indicar que esas proteínas interaccionan y que los residuos que co-evolucionan están próximos en la interfaz de interacción. De esta manera, la co-evolución tiene el potencial de proporcionar una información muy detallada de los modos de interacción entre proteínas.

Esta línea de trabajo orientada a la detección de co-evolución entre posiciones de proteínas que interaccionan se ha venido desarrollando desde los años noventa. De hecho, fue iniciada por una serie de trabajos pioneros en nuestro laboratorio, en los que se demostraba la capacidad de estas señales co-evolutivas para predecir interacciones entre proteínas (Pazos *et al.*, 1997; Pazos & Valencia, 2002). En aquel momento los métodos y la cantidad de secuencias disponibles no permitieron utilizar dicha señal para la predicción sistemática de regiones de interacción, aunque sí permitió la acumulación de evidencias indirectas que pudieron utilizarse para ayudar a desentrañar la forma de interacción en casos concretos (Filizola *et al.*, 2002).

Llegados a este punto, cabe señalar que la detección de co-evolución entre pares de residuos de proteínas diferentes que interaccionan también tiene ciertas limitaciones importantes. Entre ellas se encuentra una mayor dificultad para encontrar casos con suficientes secuencias. La aplicación de esta metodología exige establecer multitud de pares de proteínas que interaccionen, pertenecientes al mismo par de familias de proteínas. Estos emparejamientos son necesarios para observar los patrones de sustitución coordinados entre las proteínas que interaccionan. Obtener estos emparejamientos resulta problemático porque las familias de proteínas con más miembros incluyen muchos parálogos, cuya especificidad de interacción es difícil de establecer. Por tanto es preciso recurrir a información externa para construir los pares de secuencias que interaccionan. La estrategia más utilizada saca partido de la organización genómica de las especies procariotas, las cuales suelen agrupar genes asociados funcionalmente con el fin de coordinar su expresión (Jacob *et al.*, 1960; Lawrence, 2002). De esta forma los pares de proteínas que interaccionan tenderán a aparecer

próximos en sus genomas, lo que nos permitirá detectar un subconjunto de pares que muy probablemente están interaccionando. Como es evidente, esta estrategia limita el uso de la información co-evolutiva a aquellos casos de interacciones en procariotas donde se puedan recuperar suficientes homólogos de pares de proteínas que interaccionan.

A pesar de estas limitaciones, el ámbito de aplicación de la co-evolución a la predicción de regiones y modos de interacción es amplio y muy prometedor. De hecho varios trabajos recientes han mostrado excelentes resultados en la reconstrucción de decenas de complejos de proteínas (Hopf *et al.*, 2014; Ovchinnikov *et al.*, 2014).

1.2.5. Co-evolución de residuos en grupo

Hasta este momento, la co-evolución se ha discutido como un fenómeno que ocurre entre pares de residuos. Esta asunción era evidente en los métodos originales que consideraban cada par de residuos como totalmente independiente del resto (Korber *et al.*, 1993; Göbel *et al.*, 1994). Sin embargo también existe en el caso de los métodos más recientes que establecen modelos de co-dependencia de todos los residuos estudiados, ya que estos métodos pretenden recuperar una aproximación de la red de co-dependencias directas entre pares de residuos que, se asume, habría producido la matriz de co-variaciones aparentes observables. La mejora en la capacidad predictiva de estos modelos muestra que son mejores representaciones de la co-evolución entre pares de residuos. Sin embargo estas aproximaciones no consideran la posible co-evolución de residuos en grupo (señales de co-evolución compartidas por varios residuos).

El concepto de la co-evolución en grupo aún no se explorado en profundidad. Sin embargo en las últimas décadas se han estudiado extensamente grupos de posiciones que mutan coordinadamente en el contexto de la divergencia entre subfamilias de proteínas (Casari *et al.*, 1995; Lichtarge *et al.*, 1996; del Sol Mesa *et al.*, 2003; Lichtarge *et al.*, 2003; Kalinina *et al.*, 2004; Mihalek *et al.*, 2004; Reva *et al.*, 2007; Rausell *et al.*, 2010). Estas posiciones se bautizaron originalmente como *treedeterminants*, aunque finalmente se ha impuesto la denominación de posiciones determinantes de especificidad o *SDPs* (del inglés *Specificity Determining Positions*).

Es de justicia aclarar que estos grupos de residuos no se han estudiado en calidad de grupos de residuos que co-evolucionan, ya que sólo recientemente se ha propuesto esta posibilidad (Juan *et al.*, 2013). Lo que ha hecho de los *SDPs* un objeto de estudio atractivo ha sido su implicación en el establecimiento de las diferencias funcionales entre subfamilias de proteínas. Para entender la base de esta implicación será útil explicar un modelo arquetípico del proceso de aparición y establecimiento de subfamilias dentro de una familia de proteínas.

Según este modelo arquetípico las subfamilias de proteínas se fundarían por fenómenos de duplicación génica. Como parte del proceso de estabilización de la duplicación en la población o posteriormente a él, las dos copias resultantes de la duplicación experimentarían un período de divergencia entre ellas (en el que puede darse selección positiva o no). Durante esta etapa se exploran y establecen los diferentes roles funcionales de los parálogos recientes (que pueden ser funciones nuevas o formas especializadas de la función ancestral). Una vez ambas copias han adoptado un papel diferenciado y relevante, éstas se verán sometidas a selección negativa conservándose en la sucesión de especiaciones posteriores. De esta forma, en un caso ideal cada subfamilia acaba compuesta por proteínas de especies diferentes, de secuencia y función semejante, claramente diferenciadas de las secuencias y funciones de otras subfamilias

en las mismas especies. Los detalles de las dinámicas de estos procesos de duplicación son complejos, pero las duplicaciones que perduran suelen dar lugar a subfamilias funcionalmente diferentes. Siendo por tanto una fuente esencial de innovación evolutiva (Ohno, 1970; Force *et al.*, 1999; Stoltzfus, 1999; Lynch & Conery, 2000; Kondrashov *et al.*, 2002; Innan & Kondrashov, 2010).

De acuerdo con este escenario idealizado los *SDPs* cambiarían durante la fase de exploración de nuevas funciones para acabar fijándose en aminoácidos diferentes en ambas subfamilias, que aparecerán conservados en los miembros de la subfamilia por ser importante para la capacidad de estas proteínas para desarrollar la nueva función. Así los *SDPs* corresponderían a posiciones sobre las que se dan fuertes cambios de presión de selección durante este proceso (de selección neutral o negativa en el ancestro, a neutral o positiva después de la duplicación y a negativa al estabilizarse las nuevas funciones).

En su definición más estricta los *SDPs* se definen como aquellas posiciones diferentes entre subfamilias pero conservadas dentro de cada subfamilia. Esta definición subraya el hecho de que la aparición de cambios coordinados es una característica común en los grupos de *SDPs*. Esta coordinación está asociada a la conservación de la relevancia de este grupo de residuos en la función de la proteína, pero también a su capacidad para acomodar los cambios funcionales establecidos. En esta situación no sería correcto asumir que las mutaciones coordinadas observadas en *SDPs* son consecuencia de relaciones de dependencia a pares, sino que corresponden a una dependencia en grupo de posiciones que establecen la especificidad funcional de las subfamilias.

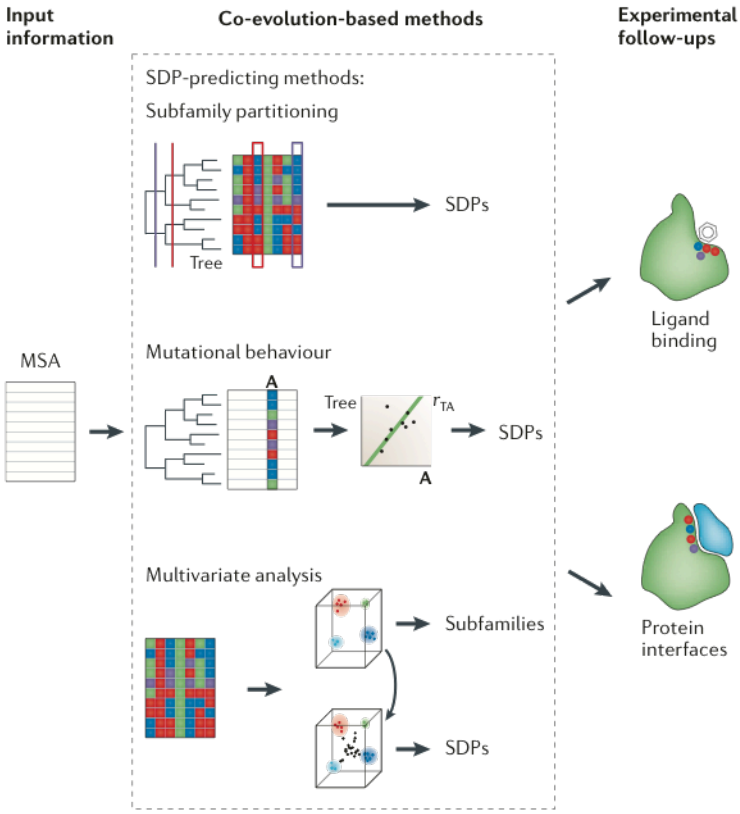


Figura 3. Implicación de residuos que co-evolucionan en grupo en especificidad funcional (adaptada de (Juan *et al.*, 2013)). Se representan tres aproximaciones diferentes para la detección de *SDPs*, que predicen residuos implicados en la especificidad de unión a ligandos y a proteínas.

Durante las dos últimas décadas se han desarrollado diferentes aproximaciones para detectar *SDPs* a partir de alineamientos múltiples. Estas aproximaciones se pueden clasificar en tres tipos (del Sol Mesa *et al.*, 2003): las que usan árboles filogenéticos (Lichtarge *et al.*, 1996; del Sol Mesa *et al.*, 2003; Lichtarge *et al.*, 2003; Kalinina *et al.*, 2004), las que buscan posiciones representativas de la variabilidad global (del Sol Mesa *et al.*, 2003; Reva *et al.*, 2007) y las que utilizan análisis multidimensionales (Casari *et al.*, 1995; del Sol Mesa *et al.*, 2003; Rausell *et al.*, 2010) (ver Figura 3).

1.2.6. Los *SDPs* son residuos funcionalmente importantes

Desde su definición original (Casari *et al.*, 1995) se ha demostrado una clara asociación entre *SDPs* y las características funcionales específicas de cada subfamilia. Esta implicación se ha determinado de forma sistemática mediante la demostración de que estos residuos suelen hallarse en sitios de unión a ligandos pequeños (del Sol Mesa *et al.*, 2003; Rausell *et al.*, 2010) o en interfaces entre proteínas (Rausell *et al.*, 2010). Igualmente se ha observado que estos *SDPs* en ocasiones también pueden reflejar cadenas de contactos implicados en la transmisión de la señal alostérica (Rodriguez *et al.*, 2010).

Los *SDPs* se han utilizado en decenas de casos para predecir sitios funcionales y/o de interacción (Hernanz-Falcón *et al.*, 2004; Oparina *et al.*, 2005; Rojas *et al.*, 2012; Peterson *et al.*, 2015), para establecer el efecto de mutaciones sobre la función de una proteína (Izarzugaza *et al.*, 2011; 2012; Katsonis & Lichtarge, 2014; Osman *et al.*, 2015) o para alterar la especificidad funcional de miembros de una familia de proteínas mimetizando la función de otra subfamilia (Bauer *et al.*, 1999; Morillas *et al.*, 2003; Shenoy *et al.*, 2006).

Estos trabajos demuestran la capacidad predictiva de estos *SDPs* y suponen un acicate para ampliar nuestra comprensión de la co-evolución al nivel de grupos de residuos de proteínas.

1.3. Co-evolución a Nivel de Proteínas

La transición conceptual de detectar co-evolución entre pares a hacerlo en grupos de posiciones puede completarse hasta llegar a considerar la co-evolución entre proteínas completas. Esta co-evolución entre proteínas también apunta a interacciones entre los agentes que co-evolucionan, es decir a interacciones entre las proteínas. Aunque, como se verá, la interpretación de la detección de señales co-evolutivas es más compleja que en el nivel de residuos.

En el 2000 Goh *et al.* (Goh *et al.*, 2000) propusieron que la co-evolución entre proteínas que interaccionan se podía detectar usando las similitudes entre los árboles filogenéticos de las familias de esas proteínas. Esta propuesta, que Goh *et al.* evaluaron para dos pares de familias de proteínas, se podría considerar el equivalente a nivel de proteínas de la Regla de Fahrenholz discutida arriba a nivel de especies. Sin embargo, en realidad vino inspirada por la co-variación detectada a nivel de residuos, lo que hace aún más evidentes los paralelismos conceptuales y metodológicos del estudio de la co-evolución a distintos niveles.

En el trabajo de Goh *et al.*, los autores emparejaron la proteínas que interaccionaban de las familias estudiadas para evaluar la correlación en los grados de divergencia de ambas familias. De hecho sus ejemplos incluían emparejamientos que involucraban a proteínas parálogas (homólogos de distintas subfamilias) e incluyeron varias veces la

misma proteína de una familia, emparejándola con todas las proteínas con que se sabía que interaccionaba. Pese al valor de esta idea, en la forma propuesta, no permitía (ni pretendía) predecir interacciones usando la co-evolución, sino detectar co-evolución entre proteínas de interacción conocida.

De hecho el uso de la co-evolución para la predicción de interacciones entre proteínas, presentaba una situación aparentemente paradójica, ya que era necesario conocer la interacción para evaluar la co-evolución. Como se ha visto en el caso de la detección de co-evolución entre residuos de proteínas diferentes, el emparejamiento de suficientes secuencias que interaccionan es un problema de difícil solución. Ahora bien, si ni siquiera se sabe si hay proteínas que interaccionan (es lo que se pretende predecir), ¿cómo realizar los emparejamientos y evaluar su grado de co-evolución?

1.3.1. Predicción de interacciones entre proteínas

Esta situación fue resuelta por nuestro laboratorio mediante el uso de relaciones de ortología biunívocas (Pazos & Valencia, 2001). Es decir, se detecta la secuencia en cada especie (sólo una) que más probablemente sea la ortóloga equivalente funcional de cada una de las dos proteínas para las que se pretendía evaluar su co-evolución. Al disponer de una única secuencia ortóloga por especie, las secuencias de ambas familias pueden emparejarse según su especie permitiendo evaluar la correlación de la divergencia entre pares de proteínas sin interacción conocida entre ellas.

Este aproximación se denominó *MirrorTree* (*MT*) (ver Figura 4) y permitió por primera vez evaluar la co-evolución entre proteínas a gran escala, incluyendo todos los posibles pares de proteínas del proteoma de una especie, *Escherichia coli* (Pazos & Valencia, 2001). Estos análisis demostraron que las correlaciones elevadas en las divergencias entre secuencias ortólogas de proteínas diferentes contenían información significativa sobre las interacciones entre proteínas. Sin embargo, al igual que ocurrió con los primeros métodos a nivel de residuos, la asociación entre la señal recuperada y las interacciones no fue lo bastante clara para realizar predicciones automáticas muy fiables. No obstante, esta metodología ha resultado muy útil en una variedad de casos concretos (McPartland *et al.*, 2007; Tiwary *et al.*, 2009; Edgar *et al.*, 2012).

Entre los desarrollos metodológicos introducidos sobre esta aproximación cabe destacar la denominada *Tol-MirrorTree* (*TMT*; del inglés *Tree of life – MirrorTree,* (Pazos *et al.*, 2005)). La idea básica de *TMT* fue desarrollada de forma independiente por nuestro laboratorio (Pazos *et al.*, 2005) y por el laboratorio del Dr. Kanehisa (Sato *et al.*, 2005). En ambos trabajos se presentaron formas alternativas de resolver una fuente importante de ruido en los análisis con *MT*: la propensión de todos los árboles de proteínas a reflejar en mayor o menor medida la historia evolutiva de las especies (el árbol de la vida). Es decir, además de la "personalidad evolutiva propia" de la que se habló con anterioridad, cada familia de proteínas también tiene una "personalidad evolutiva común" que refleja los procesos de especiación y divergencia entre las especies de las que son parte. En realidad es obvio que proteínas ortólogas de especies lejanas tenderán a ser más diferentes que aquellas de especies cercanas. Al mismo tiempo este sesgo de fondo se verá más o menos contradicho por la evolución de la proteína, cuya expresión más extrema es la transferencia horizontal. En casos de transferencia horizontal la "proteína inmigrante" no comparte la historia de la especie a la que se incorpora y su divergencia con sus ortólogos reflejará al menos parcialmente la historia evolutiva de su especie de origen.

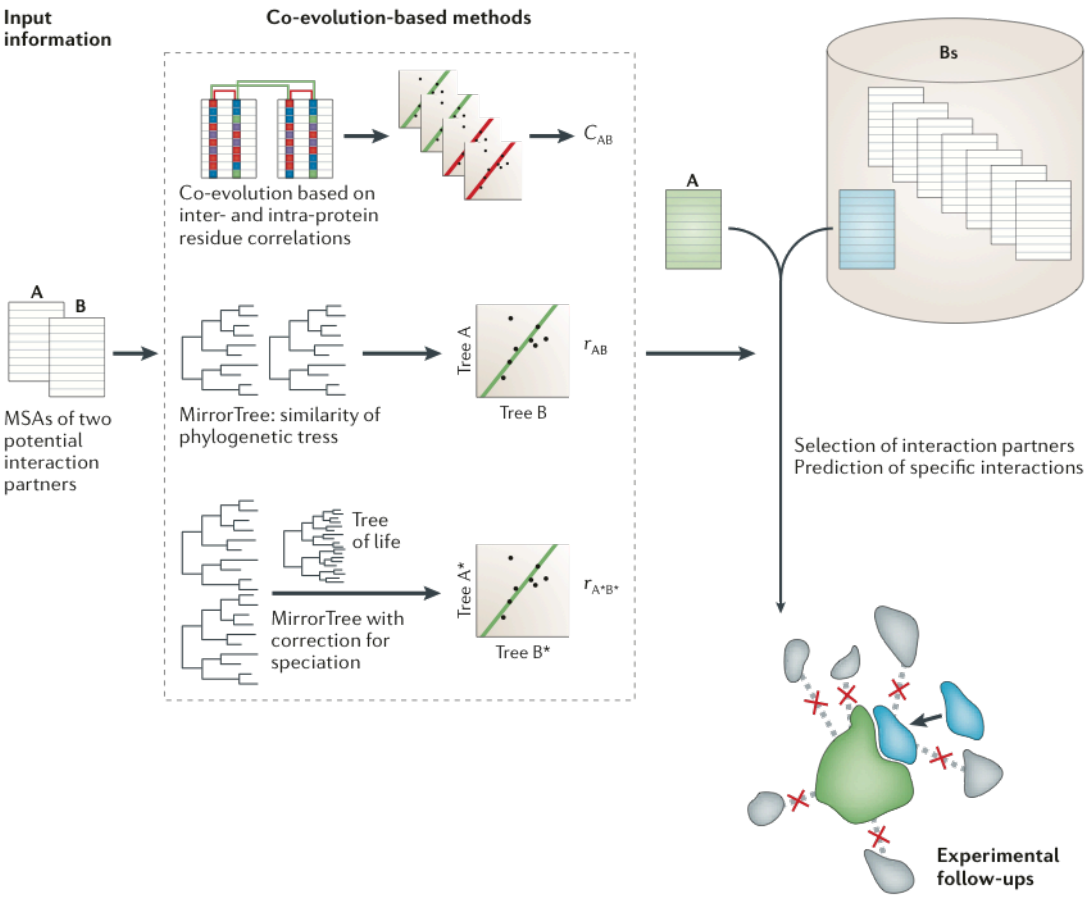


Figura 4. Predicción de interacciones basada en co-evolución entre proteínas (adaptada de (Juan *et al.*, 2013)). Se representan tres aproximaciones diferentes para la detección de co-evolución entre proteínas que se utiliza para predecir interacciones físicas o funcionales. La combinación estas predicciones con otras fuentes de información permite seleccionar compañeros de interacción con fiabilidad suficiente para ser objeto de validación experimental.

Explicaré brevemente la aproximación desarrollada por nuestro grupo con la intención de ilustrar la idea que subyace a estas y otras metodologías semejantes (ver Figura 4). *TMT* parte de un árbol de referencia que refleja la historia evolutiva de las especies (obtenido usando secuencias de 16S rRNA). Este árbol es convenientemente normalizado para reflejar grados de divergencia entre proteínas. A continuación estas divergencias son restadas de las matrices de divergencias de cada proteína del proteoma de referencia (en este caso *E. coli*). Posteriormente se establece la correlación entre estas matrices de distancias corregidas, que de esta manera pasa a reflejar similitudes asociadas a las particularidades evolutivas de cada proteína. Esta aproximación demostró ser claramente más eficaz para la detección de interacciones entre proteínas que co-evolucionan, aunque el número de falsos positivos (similitudes altas entre proteínas que no interaccionan) seguía siendo elevado (Pazos *et al.*, 2005).

Este sesgo de fondo proveniente de la historia evolutiva de las especies, es el equivalente a nivel de proteínas del sesgo discutido a nivel de residuos debido a la no independencia de las secuencias. En ambos casos existe un sesgo común a los diferentes residuos o proteínas que aumenta artificialmente la similitud entre la evolución de todos ellos. Éste es otro ejemplo de los paralelismos existentes entre los distintos niveles evolutivos y de cómo muchos conceptos y problemas son trasladables entre ellos.

Es en este punto en el que se encontraba la predicción de interacciones entre proteínas basadas en co-evolución al comienzo del trabajo relatado en esta memoria. De hecho, como quedará patente a lo largo de la exposición de resultados y discusión, el trabajo realizado incluye un desarrollo metodológico conceptualmente relacionado con las metodologías más recientes para el estudio de la co-evolución entre residuos aunque parcialmente las precedió en el tiempo.

Otro punto importante es que, hasta ese momento, la co-evolución entre proteínas se había relacionado principalmente con la interacción física entre proteínas. Sin embargo, como se ha explicado anteriormente, la co-evolución implica co-dependencia entre los agentes evolutivos, no qué tipo de interacción da lugar a dicha co-dependencia. En el caso de residuos de proteínas se trata con un sistema espacial y temporalmente compacto. Es decir todos los residuos están presentes simultáneamente en una misma estructura generalmente compacta. Por lo tanto, parece razonable que la mayoría de las co-dependencias que se den en ese contexto impliquen interacciones físicas, aunque esto deje lugar a la exploración de otro tipo de relaciones (como las asociadas a las dinámicas de plegamiento). En el caso de las interacciones entre proteínas el escenario es más complejo, ya que las relaciones de co-dependencia pueden ser mucho más variadas. Éstas incluyen aquellas esperables entre miembros de un complejo que no están en contacto directo o entre proteínas que forman parte de las mismas rutas metabólicas, entre otras. En este contexto, con asociaciones en distintos momentos y lugares, involucrando diferentes tipos de moléculas (como ligandos pequeños, ADN, ARN, etc.) y muy especialmente con la variedad de procesos en los que varias proteínas pueden influir sin interaccionar entre ellas, no sería sorprendente que una buena parte de la co-evolución entre proteínas estuviera asociada a interdependencias que no implican interacción física, pero sí funcional. Éste ha sido uno de los puntos explorados utilizando los métodos desarrollados como parte de esta tesis, y será descrito en la sección de Resultados**.**

1.3.2. Predicción de especificidad de interacción entre pares de grupos de parálogos.

Como se ha explicado, las aproximaciones tipo *MirrorTree* no precisan de ningún conocimiento previo sobre la interacción entre las proteínas analizadas, prediciendo esta interacción mediante las relaciones evolutivas entre ortólogos de dichas proteínas. En este escenario los alineamientos sólo contienen ortólogos y la interacción detectable debe estar conservada a lo largo de dichos ortólogos. Sin embargo, como también se ha mencionado anteriormente, la historia evolutiva de las familias de proteínas es más compleja e incluye eventos de duplicación génica que derivan en la aparición de subfamilias, cuyos miembros son parálogos de los de otras subfamilias.

Los métodos tipo *MirrorTree* tratan de minimizar el efecto de esta situación, por lo que no recuperan ninguna información sobre posibles especificidades de interacción diferentes entre distintos pares de subfamilias. Con la intención de afrontar este problema han surgido una serie de métodos que explotan la señal co-evolutiva para predecir estas especificidades de interacción (Gertz *et al.*, 2003; Ramani & Marcotte,

2003; Jothi *et al.*, 2005; Izarzugaza *et al.*, 2006; 2011). La situación de partida de estos métodos es la existencia de dos familias de proteínas cuyos miembros interaccionan, pero en las que se desconocen qué proteínas de una familia interaccionan con qué proteínas de la otra.

La solución planteada independientemente por dos grupos diferentes (Gertz *et al.*, 2003; Ramani & Marcotte, 2003) propone que el emparejamiento adecuado de proteínas que interaccionan debería reflejar la similitud de las historias de ambas familias. Así el problema a resolver es detectar los emparejamientos de proteínas que maximicen el parecido entre los árboles filogenéticos de ambas familias. Estos primeros métodos plantearon dos estrategias heurísticas diferentes para explorar este espacio de posibilidades. Para ello, al igual que en los métodos tipo *MirrorTree*, la similitud evolutiva se establecía comparando la divergencias evolutivas entre las proteínas de cada familia.

Estos métodos fueron evaluados sobre algunos casos de familias relativamente grandes en los que mostraron un buen comportamiento. Sin embargo, presentaban la limitación de que ambas familias debían tener el mismo número de secuencias. Además no permitían incluir información que pudiera asistir en el proceso, como establecer un grupo de pares conocidos o exigir que los emparejamientos se dieran entre proteínas de las mismas especies.

Con la intención de eliminar estas limitaciones, nuestro laboratorio desarrolló dos métodos (Izarzugaza *et al.*, 2006; 2008). En el segundo de ellos, *TAG-TSEMA* (Izarzugaza *et al.*, 2008), se estableció un sistema de etiquetas que permite definir el sub-espacio de búsqueda de cada proteína, simplificando el problema computacional y permitiendo la intervención del usuario en la definición del análisis. Este método fue evaluado no sólo en algunos casos explorados en detalle, sino también a gran escala, en un conjunto artificial de pares de dominios provenientes de las mismas proteínas. Posteriormente otros métodos han introducido mejoras orientadas a reducir el espacio de búsqueda sacando ventaja de la propia estructura topológica de los árboles (Jothi *et al.*, 2005; Hajirasouliha *et al.*, 2012), lo que si bien ayuda en la definición del problema, también introduce la necesidad de utilizar árboles filogenéticos más fiables.

1.4. Trabajo específicamente realizado en esta tesis

Durante el desarrollo de esta tesis he realizado diferentes contribuciones a algunos de los campos de aplicación de la señal co-evolutiva en proteínas. Así he contribuido al desarrollo de metodologías orientadas a mejorar la predicción de sitios funcionales y regiones de interacción entre proteínas (Tress *et al.*, 2005; Rausell *et al.*, 2010), y de interacciones entre proteínas (Pazos *et al.*, 2005; Izarzugaza *et al.*, 2006; 2008; Juan *et al.*, 2008b; García-Jiménez *et al.*, 2010; Herman *et al.*, 2011; Ochoa *et al.*, 2013; 2015). Además he aplicado metodologías basadas en co-evolución entre residuos para guiar la resolución de modelos estructurales de interacción (Hernanz-Falcón *et al.*, 2004; Juan *et al.*, 2005; Tress *et al.*, 2005). Finalmente, el desarrollo de estas metodologías ha venido acompañado por un esfuerzo por establecer recursos que permitieran a la comunidad científica su utilización, así como el acceso a los resultados (Carro *et al.*, 2006; Andrés-León *et al.*, 2009). Así mismo, como parte de este esfuerzo, he contribuido a la publicación de dos revisiones del campo (Juan *et al.*, 2008a; 2013). El conjunto de artículos y revisiones publicadas en el desarrollo de esta tesis se incluyen en el Anexo, al final de esta memoria.

Entre las contribuciones mencionadas, en esta memoria he optado por incluir únicamente el trabajo directamente relacionado con los métodos basados en co-evolución para la predicción de interacciones funcionales entre proteínas (que corresponden a (Juan *et al.*, 2008b) y a un manuscrito en preparación). De esta forma, expongo las aportaciones conceptuales y metodológicas de dos métodos que he desarrollado en este tema específico, así como sus ventajas comparativas con metodologías relacionadas. Por último incluyo los resultados obtenidos en diferentes especies bacterianas que ilustran las posibilidades de aplicación de las metodologías computacionales que he desarrollado.

2. Hipótesis y objetivos

El **objetivo principal** de esta tesis es desarrollar nuevos métodos computacionales para la predicción de interacciones funcionales entre proteínas basados en co-evolución, que permitan discriminar las señales informativas de dicha co-evolución mediante el análisis de proteomas completos. Para ello se han definido los siguientes **objetivos específicos**:

1. Desarrollar estrategias computacionales que permitan reducir el número de similitudes altas entre árboles asociadas a efectos no funcionales.

2. Desarrollar una primera metodología capaz de detectar señales de similitud entre árboles potencialmente asociadas a la co-evolución de proteínas en grupos.

3. Evaluar la calidad de las predicciones obtenidas con esta primera metodología para un organismo modelo, utilizando información externa.

4. Analizar los resultados obtenidos por esta primera metodología para explorar la relevancia de la co-evolución en grupo.

5. Desarrollar una segunda metodología capaz de detectar señales de similitud específica de pares de árboles potencialmente asociadas a la co-evolución entre pares de proteínas.

6. Evaluar la calidad de las predicciones obtenidas con esta segunda metodología en distintas especies, utilizando información externa.

7. Analizar los resultados obtenidos por ambas metodologías para ahondar en la comprensión de la presencia de fenómenos co-evolutivos en diferentes sistemas moleculares.

3. Resultados

En esta sección se presentará el desarrollo dos nuevas metodologías para detectar señales de co-evolución entre pares de proteínas desarrolladas durante el transcurso de esta tesis. Estos métodos se encuadran en el marco de la utilización de la similitud entre árboles filogenéticos de proteínas como indicativo de su co-evolución a lo largo de su historia evolutiva (ver Introducción). La primera de estas metodologías se ha denominado *ContextMirror* (*CM*, ver Figura 5), ya que utiliza las historias evolutivas de todo el proteoma como contexto para depurar e interpretar la similitud de historias evolutivas. Así mismo, la segunda metodología se ha bautizado como *ContextMirror Global* (*CMG*) por suponer el desarrollo, hasta sus últimas consecuencias, del paradigma de la detección global de las co-dependencias evolutivas entre todas las proteínas del proteoma.

Como se ha explicado en la Introducción, *TMT* (Pazos *et al.*, 2005) ya supuso un avance conceptual sobre el análisis de similitudes entre árboles de proteínas (Pazos & Valencia, 2001), al considerar que parte de esta similitud se debía al efecto de compartir la historia evolutiva de las especies en las que se encontraban. En cierto modo, *CM* y *CMG* extienden este paradigma al considerar que la historia evolutiva inferida para una proteína es una composición de distintas influencias, incluyendo el árbol filogenético de las especies, pero también (de forma esencial) de su co-evolución con una o más proteínas, de los eventos de transferencia horizontal e incluso de posibles errores metodológicos (ver Introducción).

Como consecuencia, para comprender estas metodologías es importante entender primero cuales son los efectos del sesgo filogenético en *MT*. En los artículos publicados sobre el tema (Pazos *et al.*, 2005; Sato *et al.*, 2005) se suele hacer hincapié en la tendencia a aumentar los valores de correlación entre árboles (los dos árboles representan también el efecto común de la divergencia entre las especies). Tendencia que hace más difícil la detección de señales de similitud más sutiles. Esto podría dar la impresión de que dicha tendencia es coherente a lo largo de diferentes distribuciones de especies. Es decir, que el número e identidad concreta de las especies no es relevante para esta influencia. Sin embargo, los sesgos introducidos por las diferentes especies van a depender de las distancias evolutivas entre ellas. En particular, en algunos casos, las distribuciones de distancias comparadas entre dos árboles pueden no ser unimodales. Por ejemplo, esto ocurre cuando dichas distancias provienen de un conjunto de especies con una fuerte estructura de grupos. Es decir, que incluye grupos muy divergentes compuestos por especies muy próximas. En estas situaciones la correlación de Pearson no proporciona una buena estimación de la relación entre las mismas.

Para minimizar este problema, se debe empezar por seleccionar un conjunto de especies de partida que no combine distancias evolutivas muy grandes con otras muy pequeñas. Este punto se resolvió de forma sencilla en el análisis realizado con *CM* (ver Materiales y Métodos) y de una forma algo más elaborada en el realizado con *CMG* (ver Materiales y Métodos). En el periodo transcurrido entre ambos análisis, se estudió de forma sistemática el efecto de la selección de especies (Herman *et al.*, 2011). Las conclusiones de dicho trabajo incluyen que la selección idónea de especies debe consistir en grupos de especies con distancias evolutivas semejantes entre ellas. Lamentablemente, una selección adecuada de especies no es suficiente para resolver este problema, ya que, en última instancia, las distancias que se comparan en *MT*

provienen de los pares de especies en los que se ha podido detectar ortólogos para ambas proteínas. Por lo tanto, en un análisis a gran escala, como los presentados en esta memoria, será inevitable que algunos pares de árboles den lugar a correlaciones más o menos distorsionadas. Como consecuencia, gran parte del esfuerzo realizado en los métodos presentados está enfocado en resolver o minimizar el efecto de estas distorsiones. El éxito o fracaso de este objetivo es determinante para la capacidad de extraer señales de similitud evolutiva realmente informativas de co-dependencias funcionales sostenidas en el tiempo.

Por otro lado, establecer directamente la calidad de la señal co-evolutiva no es posible, por carecer de estándares válidos de casos sólidamente definidos de co-evolución entre proteínas. Sin embargo, como se discute en la Introducción, la co-evolución es un fenómeno que debe aludir a la evolución de relaciones de co-dependencia funcional. Por lo tanto, se evaluará la calidad de la señal co-evolutiva detectada por nuestros métodos a través de su capacidad para predecir interacciones funcionales entre proteínas. De esta manera, se acentúa la vertiente práctica del desarrollo de estas metodologías, siendo obvio que mejores predicciones estarán generalmente asociadas a una mejor detección de la co-evolución.

La sección de resultados se ha dividido en dos sub-secciones principales. La primera de estas sub-secciones está dedicada a la presentación de *CM*, así como de sus resultados detectando co-evolución a lo largo procariotas para los grupos de ortólogos de proteínas de *Escherichia coli*, (la bacteria que ofrece más posibilidades de contrastar las predicciones con resultados experimentales). La segunda subsección presentará *CMG* y sus resultados en el análisis de señales co-evolutivas más recientes en 23 proteomas bacterianos diferentes.

3.1. Co-evolución específica de grupos de proteínas.

CM, a diferencia del resto de métodos descritos (ver Introducción), no pretende detectar la señal de co-evolución entre pares aislados de proteínas. En su lugar, *CM* pretende explorar y detectar escenarios de señal compartida por unas pocas proteínas. Para ello, debe eliminar las contribuciones de todo fenómeno claramente inespecífico, como el de la evolución de las especies. Sin embargo, *CM* no requiere de la especificación de cuáles son estos fenómenos, ni de cómo afectan a la evolución de las proteínas. En su lugar, *CM* utiliza los propios árboles de las proteínas del proteoma de referencia para determinar si el origen de la señal detectada es inespecífica o si aparece sólo en un número reducido de proteínas. En este sentido *CM*, está en un punto intermedio entre *MT*, que al considerar todas las correlaciones independientes entre sí recupera muchas señales inespecíficas (incluidas señales espurias asociadas a la estructura de la red de similitudes) y los métodos que infieren interacciones directas (ver Introducción y Sección 3.2) detectando señales específicas de par, pero sin considerar la existencia de co-dependencias en grupo.

CM se ha diseñado como un protocolo por pasos o capas que parte de la similitud entre las historias evolutivas de los componentes del proteoma de una especie de referencia (es decir de los resultados de *MT*). El objetivo de *CM* es detectar qué parte de esta similitud proviene de una dependencia evolutiva indicativa de interacción funcional. En consecuencia, *MT* (con algunas modificaciones metodológicas), supone el núcleo de *CM*. Sobre este punto de partida se ha desarrollado una primera capa a la que se ha denominado *Perfiles Co-evolutivos* (*PC*). Esta capa utiliza los resultados de *MT* a lo largo de un elevado número de pares de proteínas como marco común para establecer una medida robusta de la similitud evolutiva para cada par de proteínas. Finalmente, se

ha desarrollado una segunda capa cuyo objeto es establecer un protocolo para explorar la especificidad de cada señal co-evolutiva considerando el resto de similitudes detectadas para esas proteínas. Este protocolo fue implementado en el lenguaje de programación *C*.

Se empezará por describir *CM* para posteriormente presentar los resultados obtenidos por cada uno de los pasos involucrados en la metodología propuesta para la predicción de interacciones funcionales en el organismo modelo *Escherichia coli*.

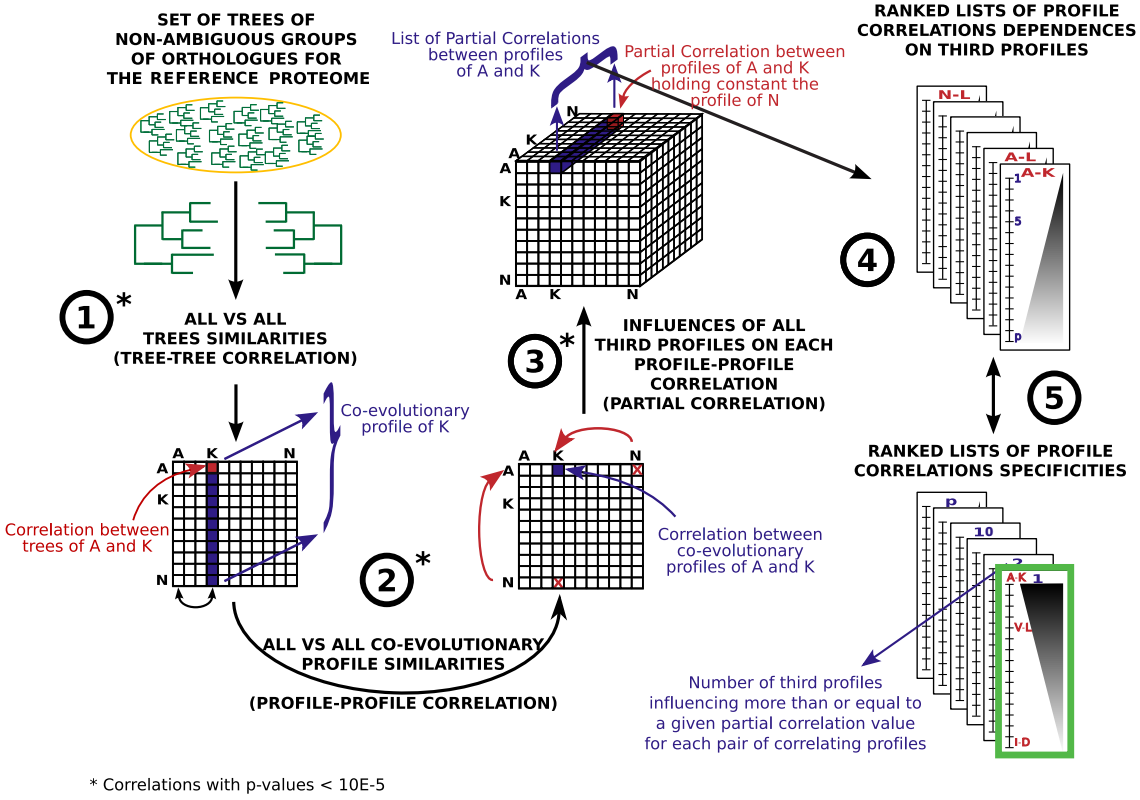


Figura 5. Esquema del protocolo de *ContextMirror* (adaptada de (Juan *et al.*, 2008b)). Se representan los cinco pasos principales del método (ver texto): **1)** *MirrorTree*; **2)** *Perfiles Co-evolutivos*; **3)** cálculo de correlaciones parciales; **4)** Ordenamiento de las correlaciones parciales para cada par de proteínas; **5)** Establecimiento de los niveles de especificidad. Los pasos 3 a 5 corresponden al paso denominado *CM* en texto principal de esta memoria.

3.1.1. *MirrorTree* (*MT*)

MT ((Pazos & Valencia, 2001), ver Figura 5.1) es una metodología desarrollada por los Doctores Florencio Pazos y Alfonso Valencia que establece las distancias evolutivas entre los miembros de un conjunto de proteínas ortólogas y las compara con las de otro conjunto de proteínas ortólogas pertenecientes al mismo grupo de organismos. Dicha comparación se establece mediante el cálculo del coeficiente de correlación de Pearson (Pearson, 1895) entre ambos vectores de distancias (r_{AB}, para los árboles de las proteínas A y B). Esta metodología se desarrolló cuando sólo existían un grupo pequeño de especies completamente secuenciadas (requisito para la definición fiable de

ortólogos), siendo su ámbito de aplicación los análisis de tamaño medio (de decenas a cientos de casos).

Es esta aproximación se emplea la correlación de Pearson para inferir la existencia de co-dependencia evolutiva entre ambas proteínas. Sin embargo, la confianza en dicha inferencia depende del número de distancias en la comparación. Como es lógico, el cálculo de correlaciones entre vectores de distancias evolutivas en *MT* requiere que dichas distancias estén representadas en ambos vectores. Sin embargo, la distribución de especies en las que es posible encontrar ortólogos para las diferentes proteínas de un proteoma es muy heterogénea. Esto supone que cada correlación entre diferentes proteínas se obtiene para un grupo diferente de distancias que involucran a diferentes especies. Los tamaños de estas comparaciones abarcan un rango desde $n = n_{min}$ hasta $n = m(m\text{-}1)/2$; donde n_{min} es número mínimo admitido de distancias evolutivas comparables entre ambas proteínas (es un parámetro establecido *a priori*) y m corresponde al número total de organismos incluidos en el análisis.

Por lo tanto, los valores de correlación obtenidos entre distintos pares de proteínas no son estrictamente comparables, ya que la inferencia de una co-dependencia evolutiva a partir de ellos lleva asociada una confianza diferente. Este problema era poco relevante en el contexto original del desarrollo de *MT,* en el que el valor acotado de m (se disponía sólo de 14 organismos completamente secuenciados y se requerían al menos 55 distancias en común, que corresponden a 11 ortólogos), suavizaba de forma importante tal limitación (a cambio de la limitación aún mayor impuesta por la carencia de suficiente información), pudiendo considerarse la correlación como un indicador de la probabilidad de interacción evolutiva.

Con la inclusión de un número mayor de especies y su aplicación a grupos grandes de proteínas se hizo preciso establecer un criterio de significación estadística. Además, en *CM*, esto es aún más necesario, ya que se pretende utilizar los valores de correlación de multitud de pares proteínas para determinar aquellos pares con una señal específica de grupos pequeños de proteínas.

Aquí, se propone una alternativa sencilla para reducir este problema estableciendo un criterio de significación estadística que permita descartar aquellos valores que podrían darse en ausencia de una correlación real. Para ello se determina un *p valor* aproximado para cada valor de correlación, establecido a partir de tablas previamente calculadas de valores críticos de coeficientes de Pearson en función del número de puntos involucrados en la comparación (n). Estas tablas se calcularon en base a los datos tabulados de la distribución *t de Student* (Student, 1908).

Estos *p valores* permiten mejorar el manejo de la heterogeneidad de los valores de n incluidos en las comparaciones entre vectores de distancias. Se estableció un nivel umbral de significancia estadística (típicamente de *p valor* $<= 10^{-5}$), de forma que sólo se consideran los valores de correlación significativos. Los *p valores* aproximados también presentan problemas importantes, ya que la distribución de los valores de distancias evolutivas dista mucho de ser normal. Esto explica la necesidad de establecer un *p valor* tan restrictivo. Como parte de esta tesis también se ha contribuido a explorar otras formas más adecuadas de establecer esta significación estadística (Ochoa *et al.*, 2015).

Es importante tener en mente que este paso de *CM* produce una matriz simétrica pero incompleta de correlaciones de Pearson entre las historias evolutivas de los pares de proteínas en el proteoma de referencia. Esta matriz contiene celdas vacías debido a la ausencia de suficientes especies en común como para realizar el cálculo o al hecho de

que la correlación obtenida no resultó ser significativa según el criterio descrito. Los valores de correlación en esta matriz siguen siendo difíciles de comparar, dada la inclusión de distintas especies en su cálculo. Reducir este problema es uno de los objetivos de los *Perfiles Co-evolutivos*.

3.1.2. *Perfiles co-evolutivos* (*PC*)

En este paso, se pretende obtener una medida más robusta de la similitud entre la historia evolutiva de dos proteínas que sea razonablemente comparable a lo largo de diferentes pares de proteínas. Para ello se utilizó la información de toda la red de similitudes entre árboles definidas por *MT*.

Por tanto, *PC* (ver Figura 5.2) parte de la matriz de valores de los coeficientes de Pearson de *MT* para el proteoma de referencia. Esta matriz tiene dimensiones p x p en la que cada columna (y cada fila) representa a una proteína. En esta matriz se define a cada columna como el perfil co-evolutivo de la correspondiente proteína (de ahí el nombre de *Perfiles Co-evolutivos*). Así este vector contiene todos los valores de los coeficientes de Pearson entre el árbol filogenético de esta proteína y los de todas las demás incluidas en el análisis.

Para detectar pares de perfiles co-evolutivos coherentes a lo largo de todas las comparaciones realizadas por *MT*, se calcula de nuevo el coeficiente de Pearson (r'_{AB}, entre todos los pares de vectores de los perfiles co-evolutivos. La significación estadística de estas correlaciones se establece de forma similar a la descrita para *MT*.

La idea básica de los *Perfiles Co-evolutivos* es que los pares de proteínas con fuertes correlaciones deben presentar unas correlaciones similares con el resto de las proteínas del proteoma, sin importar las especies involucradas en cada comparación. Éste es un criterio exigente, cuya aplicación conlleva varios efectos.

En primer lugar, este criterio reduce la posibilidad de introducir falsos positivos asociados a problemas en la distribución distancias, ya que este problema debería persistir de forma semejante a lo largo de las distintas distribuciones implicadas en todas las comparaciones con terceras proteínas. De hecho, este criterio favorece la detección de una red coherente de similitudes evolutivas válida para la mayoría de las especies en las que se detecten ortólogos de las proteínas comparadas. Por ejemplo, correlaciones asociadas a casos de transferencia horizontal tenderán a ser penalizadas, ya que éstas son mucho más dependientes de las especies incluidas en la comparación.

Además, estas comparaciones a lo largo de todo el proteoma involucran a un número elevado de valores (cientos o miles de proteínas), por lo que se reducen los problemas de heterogeneidad de los valores de correlación, que son mucho más importantes cuando se incluyen casos con pocos valores (en los que sólo para las correlaciones extremadamente elevadas se puede afirmar que son significativas). Este paso también prioriza de forma natural a aquellas correlaciones suficientemente altas para dar lugar a patrones coherentes, reforzando el filtro realizado en el paso anterior. Por todo esto, *PC* está diseñado para recuperar similitudes evolutivas fuertes que son comparables entre ellas y con la mayoría de las proteínas en el proteoma.

En consecuencia, el resultado de este paso es la transformación del conjunto de coeficientes de Pearson entre matrices de distancias evolutivas de grupos de ortólogos, a un conjunto de coeficientes de Pearson entre perfiles co-evolutivos razonablemente comparables entre sí. Es importante tener claro que la matriz resultante de este paso sigue sin ser completa, ya que siguen existiendo pares de proteínas sin valores de

correlación significativos o sin suficientes valores en común para recuperar una correlación fiable.

3.1.3. *ContextMirror (CM)*

Una vez se dispone de una matriz de dependencias evolutivas o co-evoluciones entre proteínas razonablemente comparables y de mayor calidad, se pretende explorar la influencia de terceras proteínas en dichas dependencias. Como se ha comentado, esta estrategia considera la existencia de similitudes evolutivas que involucran a más de dos proteínas, interfiriendo en nuestra capacidad para distinguir entre co-evoluciones con distinto grado de especificidad y similitudes evolutivas inespecíficas.

En el contexto de las correlaciones, se puede explorar la influencia en la correlación de dos variables de una o más terceras mediante el uso de las correlaciones parciales. Por ejemplo, la correlación parcial de dos perfiles co-evolutivos frente a un tercero (correlación parcial de primer orden) establece la parte de la dependencia estadística entre los dos primeros perfiles que no se explica por su mutua dependencia con el tercero. Este tipo de análisis también se pueden realizar entre un par de proteínas frente a un grupo de otras (orden = número de terceras proteínas).

En el caso de *CM*, se estableció un protocolo para la exploración de las relaciones de primer orden con el fin de detectar pares o grupos de proteínas que muestren señales de co-dependencia evolutiva específicas de ellos e independientes del resto. Esta aproximación permite lidiar de forma natural con la matriz de correlaciones obtenida, ya que no requiere que la matriz de correlaciones sea completa. En la sección 3.2 se describirá la aplicación de una estrategia alternativa que permite obtener correlaciones parciales controlando por todas las terceras proteínas.

Para calcular las correlaciones parciales de primer orden se acude a la fórmula (Guttman, 1938):

$$\rho'_{AB.X} = \left(\frac{r'_{AB} - r'_{AX} \times r'_{BX}}{\sqrt{\left(1 - r'^{2}_{AX}\right) \times \left(1 - r'^{2}_{BX}\right)}} \right),$$

donde, en este caso, $\rho'_{AB.X}$ es la correlación parcial entre los perfiles co-evolutivos de las proteínas A y B, manteniendo constante el perfil de la proteína X, y r'_{AB}, r'_{AX} y r'_{BX} son los coeficientes de Pearson entre los perfiles co-evolutivos de A y B, A y X, y B y X, respectivamente.

Se calculan todas las combinaciones de correlaciones parciales de dos perfiles co-evolutivos manteniendo constante cada tercero en nuestro conjunto (ver Figura 5.3). Esto supone un máximo de $n \times (n-1) \times (n-2)/2$ correlaciones parciales, donde n es el número de proteínas. A continuación se ordena de menor a mayor las correlaciones parciales para cada par de perfiles (ver Figura 5.4). Así, la primera correlación parcial (es decir la de menor valor) de cada una de estas listas ordenadas representa qué parte de la dependencia entre los dos perfiles no puede ser explicada por el tercer perfil que más explica de la relación. Por lo tanto, esta correlación parcial mínima es una aproximación de la intensidad de la co-evolución específica del correspondiente par de proteínas. Esta aproximación no debe confundirse con la correlación parcial controlando por todas las terceras variables, que proporciona una estimación más precisa de la señal específica del par (ver Sección 3.2).

Es importante advertir que, *CM* aborda de forma implícita la influencia en las predicciones de similitudes evolutivas debidas al árbol de las especies compartido por todas las proteínas (ver Introducción). Esta influencia se había tratado anteriormente de forma explícita por *Tol-MirrorTree* definiendo un árbol de referencia cuya influencia se eliminaba de los árboles de proteínas (Pazos *et al.*, 2005). En *CM* su influencia se elimina de forma implícita en tanto que fuente de similitudes inespecíficas y, por lo tanto, presente en un número elevado de proteínas. De hecho, *CM* resuelve de esta manera cualquier sesgo en la distribución de distancias evolutivas que resulte en la tendencia de un grupo de secuencias grande a mostrar correlaciones altas, ya que éstas están consideradas en las correlaciones parciales frente a proteínas de ese grupo.

La flexibilidad del protocolo de *CM*, también permite estudiar la posible existencia de co-evolución en grupos de proteínas (asociada a complejos moleculares, rutas metabólicas, etc.). Para ello, se relaja la exigencia de especificidad co-evolutiva definiendo diferentes niveles de especificidad de acuerdo a la posición en cada lista ordenada de correlaciones parciales (ver Figura 5.5). Por ejemplo, el nivel 1 corresponde a los valores de la primera (la menor) correlación parcial de cada lista, mientras que el nivel 10 corresponde al décimo valor de cada lista de correlaciones parciales. Cada uno de estos niveles se puede interpretar como la intensidad de la co-evolución de cada par de proteínas permitiendo la existencia de un determinado número de proteínas que contribuyan a dicha dependencia. Esto no necesariamente implica que las *n* terceras proteínas ignoradas en el nivel *n* presenten la misma señal. Sin embargo, dado el carácter global de las señales no específicas, niveles relativamente altos deberían mantener la señal de grupos pequeños con historias evolutivas muy similares al par, eliminando la señal compartida por muchas terceras proteínas (ver Figura 6).

La lógica de esta estrategia es observar si la relajación del nivel de especificidad permite rescatar relaciones co-evolutivas en grupo. Estos casos podrían quedar enmascarados en el nivel 1, ya que una parte importante de la correlación para cada par de proteínas del grupo sería explicable por su dependencia de los otros miembros del grupo. Para ilustrar este punto se explorarán dos pares concretos de proteínas de *Escherichia coli*: NuoF-NuoH y NudE-PepA (corresponden a dos pares extraídos del análisis presentado en la sección 3.1.4). NuoF-NuoH son dos miembros del complejo NADH oxidoreductasa, complejo multienzimático fundamental de la cadena respiratoria de *E. coli*. Por otro lado, NudE-PepA es un par sin relación funcional conocida, pero que tiene un valor para *CM-1* (y para *PC*) muy similar (ligeramente mayor) al de NuoF-NuoH, aunque ambos son bajos (y altos para *PC*). Sin embargo los valores de ambos pares cambian de forma radicalmente diferente con el nivel de *CM* (ver Figura 6). Mientras que NuoF-NuoH rápidamente alcanza un valor alto de ρ'(ρ'*(CM-10)* = 0,74), el par sin relación funcional NuoE-PepA sufre un ascenso mucho más paulatino y muestra valores mucho menores que NuoF-NuoH en los 400 primeros niveles. Esto muestra que la similitud en los perfiles (*PC*) detectada para el par negativo era muy inespecífica, aunque de hecho pueda ser muy similar a la del par positivo (ambos mayores de 0,95). En cambio, NuoF-NuoH, presenta una fuerte señal que no es específica del par, sino que surge al admitir la influencia de sus compañeros de complejo. Esto sugiere que los casos de co-evolución en grupo pueden quedar enmascarados tanto en niveles demasiado específicos (su compañeros de grupo reducen mucho su señal) como en los demasiado inespecíficos (las señales inespecíficas hacen que los pares que co-evolucionan y los que no, sean difíciles de distinguir).

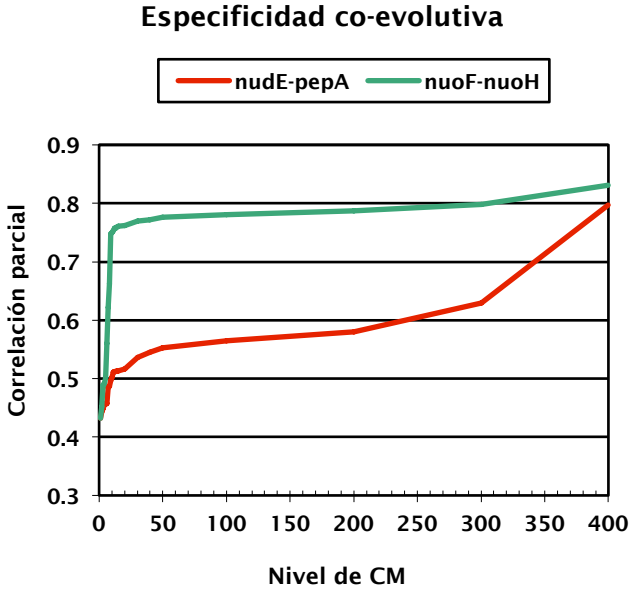


Figura 6. Correlación parcial para distintos niveles de *CM* (adaptada de (Juan *et al.*, 2008b)). Ejemplo de cómo varían los valores de la correlación parcial correspondientes a diferentes niveles de *CM*. El par de proteínas NuoF-NuoH pertenecen al mismo complejo (la NADH deshidrogenasa), pero su señal co-evolutiva no es claramente superior a la del par NudE-PepA (sin relación funcional conocida) ni para niveles muy específicos (*CM-1*) ni a niveles muy poco específicos (más allá de *CM-400*), pero si en los niveles intermedios.

3.1.4. Evaluación de *ContextMirror* y estudio de las relaciones co-evolutivas en *Escherichia coli*

La evaluación de la capacidad de una metodología como la propuesta es siempre problemática, ya que no se dispone de ninguna medida experimental directa de co-evolución entre proteínas. Sin embargo, como se discutió previamente en la Introducción, la co-evolución debe darse en pares de proteínas funcionalmente co-dependientes. De hecho, estudios previos (Pazos & Valencia, 2001; Pazos *et al.*, 2005; Sato *et al.*, 2005) encontraron una relación estadísticamente significativa entre la co-evolución, definida en términos de la similitud de árboles filogenéticos (*MirrorTree*) y la interacción física entre proteínas, lo que permitió utilizar la co-evolución para predecir interacciones.

Por lo tanto se estableció una evaluación de la capacidad predictiva de interacciones entre proteínas, tanto físicas como funcionales. Esta evaluación satisface dos objetivos de forma simultánea: ayuda a determinar si esta metodología amplía los límites (aún inciertos) de la relación entre co-evolución y relaciones funcionalmente importantes; y permite determinar la utilidad de dicha aproximación. Sin embargo, esta evaluación adolece de algunas limitaciones inevitables, ya que se desconoce totalmente cuál es el número de pares de proteínas que han co-evolucionado y qué porcentaje representan de las interacciones funcionales conocidas.

En esta situación, resulta importante la comparación de la capacidad predictiva de distintas metodologías afines conceptualmente. Dado que, como se ha explicado en las secciones previas, nuestra metodología está estructurada en tres pasos claramente

separables, se realizará una evaluación del comportamiento de cada paso para los pares de proteínas con una señal co-evolutiva más clara. Así mismo, se incluirán dos niveles de especificidad de *CM* (niveles 1 y 10), con la intención de explorar señales co-evolutivas que pudieran estar asociadas con pares o con grupos mayores de proteínas. Como ya se ha comentado el primero de esos pasos es *MirrorTree*, el cuál servirá de referencia para determinar las mejoras obtenidas.

3.1.4.A. Evaluación de la capacidad predictiva de ContextMirror

Dados los problemas comentados para establecer un conjunto razonable de pares de proteínas que deberían/podrían mostrar co-evolución, se determinará el comportamiento de las primeras predicciones de cuatro medidas diferentes de co-evolución: *MT*, *PC*, *ContextMirror* Nivel 1 (*CM-1*) y *ContextMirror* Nivel 10 (*CM-10*) en una serie de conjuntos diferentes de relaciones funcionales entre proteínas.

Entre los conjuntos de referencia utilizados se encuentran:

- Conjunto de interacciones físicas establecidas en experimentos a pequeña escala y por lo tanto muy fiables (conjunto *LT_PPI*) (Bader *et al.*, 2001; Xenarios *et al.*, 2002; Zanzoni *et al.*, 2002; Kerrien *et al.*, 2007). Este conjunto representa una infraestimación muy fuerte de las interacciones físicas en *E. coli* (refleja menos de 4.000 interacciones entre 812 proteínas).
- Diferentes conjuntos de complejos de proteínas en los que se evalúa la presencia del par de proteínas que co-evoluciona en el mismo complejo (puede o no implicar interacción física). Entre estos conjuntos se pueden distinguir:
  - Complejos bien establecidos cuya composición ha sido confirmada manualmente a partir de experimentos muy fiables realizados a pequeña escala (conjunto *LT_COMPLEX*) (Keseler, 2004). Este conjunto es una infrarrepresentación muy fuerte de los complejos en *E. coli* (contiene 245 complejos).
  - Complejos derivados de experimentos a gran escala de aislamiento de complejos con una gran cobertura de las proteínas de *E. coli*, pero de menor fiabilidad (conjunto *HT_COMPLEX*) (Butland *et al.*, 2005; Arifuzzaman, 2006).
- Conjuntos de rutas metabólicas definidas manualmente a partir del conocimiento bioquímico acumulado en organismos modelo como *E. coli* (conjuntos *ECOCYC_PWY* y *KEGG_PWY*) (Kanehisa *et al.*, 2004; Keseler, 2004). Al forma similar a los conjuntos de complejos, en este caso se evalúa la presencia en la misma ruta metabólica.

Estos conjuntos se presentan más detalladamente en la sección de Materiales y Métodos, así como la determinación de sus conjuntos complementarios de pares de proteínas que no interaccionan funcionalmente.

En esta evaluación se estudiará la precisión en los *n* primeros casos ($100 <= n <= 2000$) respecto a cada conjunto de interacciones *i* definida como:

$$Precisión_{ni} = \left(\frac{TP_{ni}}{TP_{ni} + FP_{ni}} \right),$$

donde TP_{ni} son las predicciones acertadas y FP_{ni} las erradas, para el conjunto de validación *i* entre las *n* primeras.

Esta evaluación permite observar el comportamiento de las valores elevados de co-evolución estimada por cada método. De esta manera, se puede determinar su capacidad predictiva en un contexto de predicción realista, donde sólo se consideran estos casos. Sin embargo, esta aproximación no evalúa la capacidad de recuperación del total del conjunto de pares positivos, lo cual como se ha comentado sería difícil de interpretar dado el carácter de la relación entre la co-evolución y estos conjuntos de asociaciones. Por ejemplo, ¿es esperable que todos los pares de proteínas que participan en una ruta metabólica co-evolucionen? Aún es más, ¿debería esta co-evolución darse de forma consistente a lo largo de las especies incluidas en nuestros árboles de ortólogos? Es evidente que la respuesta en ambos casos es negativa, ya que ni todas las proteínas en una ruta tienen el mismo nivel de co-dependencia funcional entre sí, ni todos procesos co-evolutivos han estado presentes en todos los linajes desde el origen de la vida. Expresado de forma sencilla, en este contexto, es más útil evaluar si los pares que co-evolucionan están relacionados funcionalmente y no tanto si los pares relacionados funcionalmente co-evolucionan.

Para evaluar las mejoras atribuibles a la nueva metodología, se utilizaron como punto de referencia los resultados del método original *MT*. La precisión de este método para los diferentes conjuntos de prueba se muestra en la Figura 7A. Conforme con lo reportado previamente (Pazos & Valencia, 2001), el enfoque sencillo de *MT* es capaz de captar cierta señal de co-evolución relacionada con las interacciones de proteínas. Sorprende el hecho de que, mientras esta relación es particularmente evidente para los complejos establecidos manualmente de *EcoCyc* (*LT_COMPLEX*) (precisión próxima al 0,4 para las 500 primeras predicciones), parece no haber relación con los complejos obtenidos en experimentos a gran escala (*HT_COMPLEX*).

Cuando se comparan frente a frente los resultados de *PC* con los obtenidos por *MT*, se observa una drástica mejoría hasta las primeras 500 predicciones. Por ejemplo, se obtiene una precisión del 100% para las 100 mejores predicciones de *PC* cuando se evalúa respecto a *LT_COMPLEX*, frente al 50% obtenido por *MT* (ver Figura 7B). Para valores superiores a las primeras 500 predicciones, la precisión (para *LT_COMPLEX*) es similar a la de *MT*. Esta mejora en la parte superior de la lista supone una reducción en el número de falsos positivos producidos previamente por *MT*. Esta tendencia se observa también en los conjuntos de rutas metabólicas (*KEGG_PWY* y *ECOCYC_PWY*), pero no así en los dos conjuntos con menos relación con *MT* y *PC* (*i.e. HT_COMPLEX* y *LT_PPI*).

En el caso de *CM-1* se observa una mejora sustancial. *CM-1* elimina una gran proporción de los falsos positivos introducidos por las tendencias evolutivas más inespecíficas, como los atribuibles al proceso de especiación (Pazos *et al.*, 2005; Sato *et al.*, 2005). De hecho, la mejoría observada en este punto fue muy grande (ver Figura 7C), duplicándose prácticamente la precisión de los pasos previos para las primeras 1.000 predicciones (de 0,31 a 0,65 para los complejos de *LT_COMPLEX*, de 0,27 a 0,49 para *KEGG_PWY*, de 0,26 a 0,43 para *ECOCY_PWY*, de 0,06 a 0,23 para *LT_PPI*), excepto para *HT_COMPLEX* que se mantiene muy bajo (0,09 para ambos).

Finalmente, en la Figura 7D se muestran los resultados de *CM-10*. Estos resultados fueron aún mejores que los obtenidos para el *CM-1*, lo que indica que algunas relaciones de valor son enmascaradas al filtrar la influencia de terceras proteínas. En particular, se observa una mejora evidente para las primeras predicciones de los conjuntos de rutas metabólicas. Por ejemplo, las 100 mejores predicciones para *KEGG_PWY* y *ECOCYC_PWY* muestran una precisión de 0,96 con *CM-10* frente a 0,78.y 0,69 respectivamente con *CM-1*.

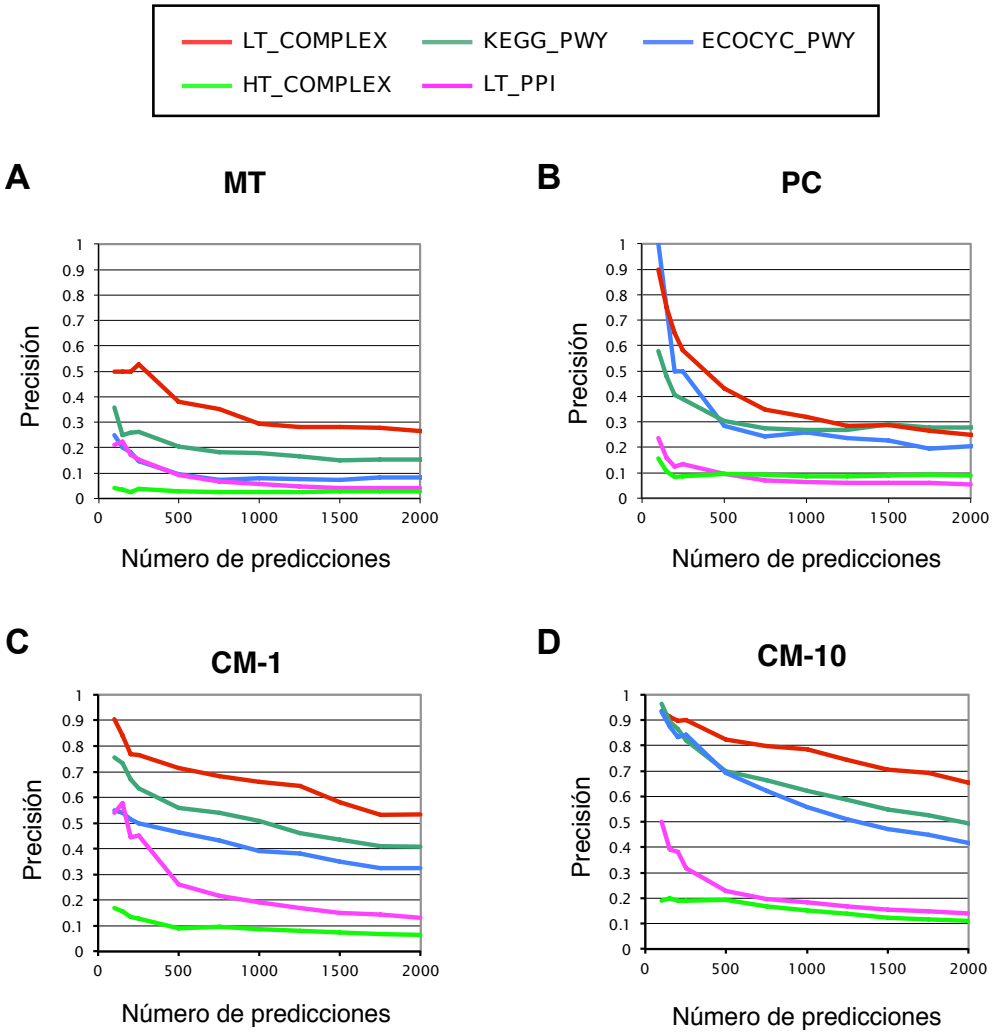


Figura 7. Precisión de los diferentes pasos de *CM* (adaptada de (Juan *et al.*, 2008b),). Precisión para las primeras predicciones de cada paso para los conjuntos de complejos bien establecidos (*LT_COMPLEX*), complejos provenientes de experimentos a gran escala (*HT_COMPLEX*), rutas metabólicas (*KEGG_PWY* y *ECOCYC_PWY*) e interacciones físicas provenientes de experimentos a pequeña escala (*LT_PPI*). **A)** *MirrorTree*. **B)** *Perfiles Co-evolutivos*. **C)** *ContextMirror* Nivel 1. **D)** *ContextMirror* Nivel 10. Para una descripción de los conjuntos ver Materiales y Métodos.

De hecho, si se analiza el cambio en la precisión con los niveles de *CM* para las 1000 primeras predicciones (ver Figura 8) se observa que *CM-10* corresponde al nivel con mejores resultados para los tres conjuntos más fiables que representan grupos funcionales (i.e. *LT_COMPLEX*, *KEGG_PWY* y *ECOCYC_PWY*). En estos tres conjuntos resulta evidente el efecto de nuestra estrategia, ya que la precisión aumenta en los 10 primeros niveles, al permitir la existencia de grupos que co-evolucionen. A partir de *CM-10*, se produce un ligero descenso en la precisión que correspondería a la inclusión de señales no co-evolutivas. Sin embargo, como cabía esperar, los resultados son muy robustos a la elección del nivel, ya que existen un gran número de proteínas que reflejan las similitudes asociadas a fenómenos globales. En cuanto a

HT_COMPLEX, sigue un patrón semejante aunque menos marcado, alcanzando su máximo a *CM-50* (0,21) y descendiendo posteriormente (0,17 para *CM-500*).

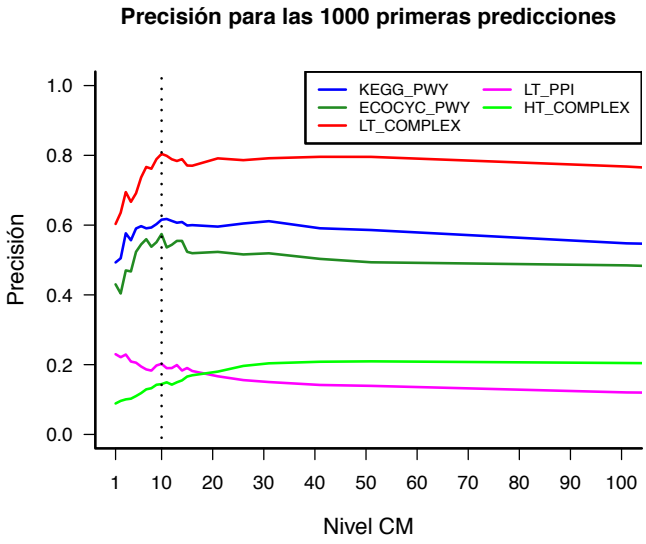


Figura 8. Precisión de diferentes niveles de *CM*. Progresión de la precisión para las primeras 1.000 predicciones a lo largo de diferentes niveles de *CM*. Los conjuntos de evaluación son los mismos descritos en la Figura 7 Los conjuntos de evaluación relacionados con interacciones funcionales en grupo (*LT_COMPLEX*, *HT_COMPLEX KEGG_PWY* y *ECOCYC_PWY*) mejoran en los primeros niveles para estabilizarse más adelante. En cambio, el grupo de interacciones físicas (*LT_PPI*) sigue el patrón inverso.

Otro caso interesante es el de *LT_PPI*, que nunca alcanza niveles elevados de precisión (un máximo de 0,23 para las 1000 primeras predicciones). *LT_PPI* no presenta una mejora en sus predicciones al relajar la condición de especificidad de la señal (al aumentar el nivel), sino que éstas empeoran. Este conjunto es el único que refleja interacciones físicas entre pares de proteínas. Estas interacciones son específicas del par y parece lógico que su señal co-evolutiva también lo sea.

Es importante señalar que aunque *CM-10* representa un máximo de precisión para los conjuntos mejor predichos, para diferentes pares de proteínas el nivel óptimo podría variar (en función de si están o no involucrados en una relación funcional que implique a más proteínas). También es de reseñar que *CM-10* produce un número mayor de predicciones que *CM-1* (19.955 frente a 2.327 predicciones con un *p valor* <= 10^{-6}).

En las cuatro estimaciones de la co-evolución presentadas, resulta evidente que la asociación de la señal de co-evolución con aquellos complejos entre proteínas definidos manualmente es la más clara y robusta de las analizadas. Por tanto, se decidió comparar la capacidad discriminativa para predecir estos complejos de *CM-10* y de la versión mejorada de *MT* utilizada en esta tesis.

Con esta intención se recurrió a un análisis de las respectivas curvas características operativas del receptor o curvas *ROC* (del inglés *Receiver Operating Characteristic*). Las curvas *ROC* son una representación gráfica de la *sensibilidad* de un clasificador frente a *1 - especificidad* a lo largo de un rango de predicciones ordenadas de más a menos informativas. Las curvas *ROC* permiten evaluar la capacidad predictiva de un

método, en función de su situación relativa respecto a la diagonal. Esta diagonal representa la predicción trivial, ya que cuando la *sensibilidad* es igual a *1 – especificidad* la capacidad discriminativa entre casos positivos (en nuestro caso pertenencia al mismo complejo de proteínas) y negativos (*no* pertenencia al mismo complejo de proteínas) es nula.

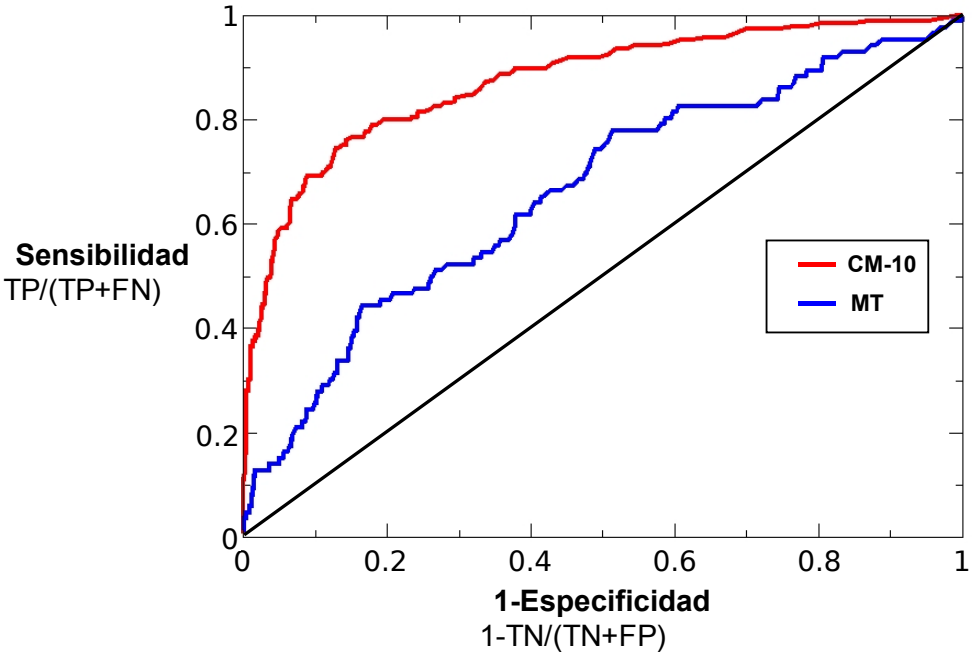


Figura 9. Curva *ROC* de *CM-10* y *MT* para complejos bien establecidos. Representación de la *sensibilidad* frente *1 - especificidad* (ver Materiales y Métodos) a lo largo de lista ordenada de las primeras 50.000 predicciones de *MT* y *CM-10*. En esta curva *CM-10* muestra una mayor capacidad de discriminación (altas sensibilidades y especificidades) que *MT*, para aquellos pares de proteínas presentes en el mismo complejo (*LT_COMPLEX*) incluidas entre las primeras 50.000 predicciones (ver texto principal).

Como se ha comentado anteriormente, la exhaustividad de estas predicciones es difícil de evaluar, por lo que en este trabajo se optó por centrarse en las primeras 50.000 predicciones de *MT* y de *CM-10* (*p valor* <= 10^{-6}). Por tanto, en este análisis, el número total de interacciones reales corresponde a aquellas presentes entre las primeras 50,000 predicciones (ver Materiales y Métodos). Este número de predicciones es muy superior a nuestras mayores estimaciones del número de posibles positivos (supondría que cada proteína co-evoluciona con más de 10 compañeras). Como se puede observar en la Figura 9, tanto *MT* como *CM-10* se encuentran claramente por encima de la diagonal, lo que muestra que en ambos casos existe una capacidad discriminativa por encima de un clasificador aleatorio. En particular, *CM-10* muestra una curva claramente por encima de la de *MT*, lo que se refleja en los valores de las áreas bajo la curva de 0.87 para *CM-10* y 0.66 para *MT* (1 corresponde a un clasificador perfecto y 0.5 a uno aleatorio). Esto

supone una mejora importante en la capacidad de detectar señales co-evolutivas informativas acerca de complejos de proteínas.

3.1.4.B. Estudio del significado funcional de los pares que co-evolucionan detectados por CM

La primera observación que se puede hacer es que las cuatro medidas de co-evolución presentan un orden de asociación consistente con los diferentes tipos de relaciones funcionales (ver Figura 7). Así, en todos los casos, se observó que los complejos establecidos manualmente a pequeña escala (*LT_COMPLEX*) están más relacionados con co-evolución que las proteínas pertenecientes a la misma ruta metabólica (*KEGG_PWY* **y** *ECOCYC_PWY*). A su vez, las proteínas de la misma ruta lo están mucho más que las interacciones físicas determinadas experimentalmente (*LT_PPI*). En todos los casos, la relación entre co-evolución y proteínas pertenecientes a los mismos complejos según experimentos a gran escala (*HT_COMPLEX*) es prácticamente inexistente.

Así mismo, es interesante observar que son las relaciones asociadas a grupos de proteínas (complejos y rutas metabólicas) las que muestran las mejoras más evidentes a lo largo del protocolo propuesto. Aunque las interacciones físicas a pequeña escala también mejoran, este efecto queda restringido a las 500 primeras predicciones, sugiriendo que una buena parte de la señal co-evolutiva refleja relaciones funcionales en grupo. Este efecto puede venir dado por la incapacidad de *CM* de recuperar las señales co-evolutivas más específicas (ver sección 3.2). A esto, se debe añadir la escasez de estos datos, ya que las interacciones físicas a pequeña escala se estudian de forma individual. Por lo tanto, asumir que se conocen todas las interacciones de una proteína presente en el conjunto (ver Materiales y Métodos) es una asunción más arriesgada que en el caso de los complejos y las rutas metabólicas.

Entre estos grupos de relaciones funcionales destaca el contraste entre el comportamiento de los dos conjuntos de complejos de proteínas. Mientras los establecidos manualmente son siempre los mejor predichos, los establecidos mediante técnicas experimentales de aislamiento de complejos a gran escala casi no parecen presentar ninguna relación.

Es bien conocido que muchas metodologías experimentales a gran escala tienden a recuperar un elevado número de falsos positivos (Mering *et al.*, 2002). Éste podría ser el caso en nuestro conjunto *HT_COMPLEX*. Para explorar este punto se empezó por establecer el nivel de concordancia de los complejos a gran escala con los complejos a pequeña escala y con las rutas metabólicas. En ambos casos, *HT_COMPLEX* tiene una capacidad casi nula de predecir a los otros conjuntos (precisiones menores de 0,15, Figura 10A y Figura 10B).

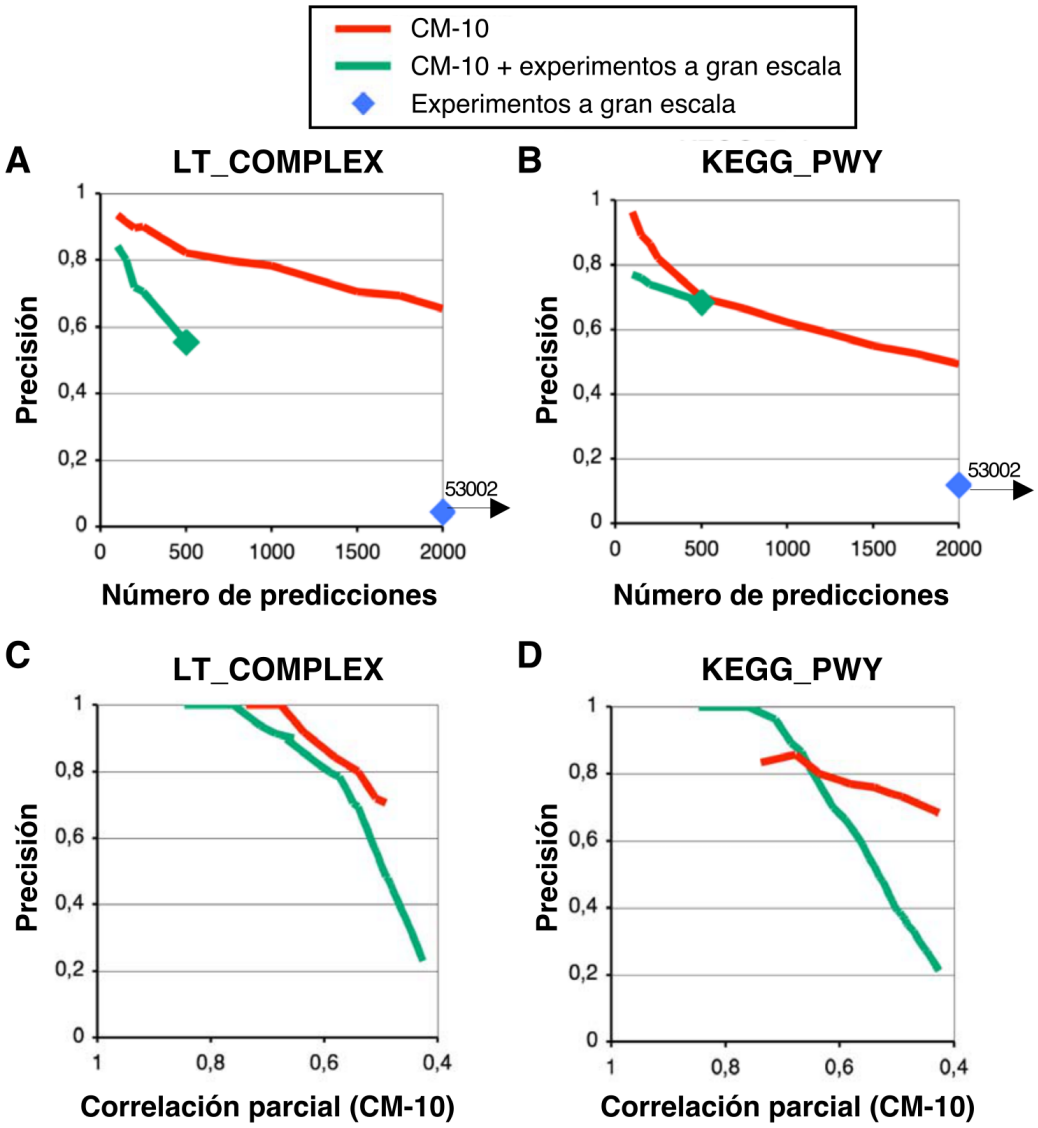


Figura 10. Precisión de la combinación de *CM-10* con complejos determinados a gran escala. **A)** Precisión para de los *CM-10*, los pares provenientes de experimentos a gran escala (Butland *et al.*, 2005; Arifuzzaman, 2006) y los pares de estos experimentos predicho por *CM-10* evaluados mediante complejos bien establecidos (*LT_COMPLEX*). **B)** Lo mismo que A) evaluado mediante anotaciones de rutas metabólicas (*KEGG_PWY*). **C)** Precisión de frente a la correlación parcial de *CM-10* para todas las predicciones de *CM-10* y sólo para aquellas recuperadas en los experimentos a gran escala. Esta precisión es evaluada mediante complejos bien establecidos. **D)** Lo mismo que C) evaluado mediante rutas metabólicas. Se observa que, aunque *CM-10* tiene mejor precisión global que su combinación con los pares experimentales, dicha combinación permite rescatar pares confirmados de *CM-10* con correlaciones parciales más bajas.

Dado el bajo nivel de concordancia de los conjuntos a gran escala con las anotaciones más fiables, se decidió investigar si era posible utilizar la señal co-evolutiva para mejorar estos valores. Para ello, se determinó el valor de *CM-10* para las relaciones entre proteínas presentes en *HT_COMPLEX*. Como se puede observar en la Figura 10A

y Figura 10B los pares de este conjunto que co-evolucionan (unos 500 pares) presentan precisiones mucho mejores en la predicción, tanto frente a *LT_COMPLEX* como a *KEGG_PWY*. Estas precisiones son menores de lo observado para los valores más altos de co-evolución (ver Figura 10A y Figura 10B), pero mayores que los obtenidos por los mismos valores de co-evolución sin considerar su presencia en el conjunto a gran escala (Figura 10C y Figura 10D). De esta manera, queda claro, que combinar la señal co-evolutiva con la información de experimentos a gran escala permite aumentar la fiabilidad de las relaciones detectadas.

3.1.4.C. Exploración de casos concretos de predicciones de ContextMirror

En esta sección se utilizarán casos concretos para ilustrar el comportamiento de *ContextMirror*. En particular, se explorará la relación entre especificidad de la señal co-evolutiva (es decir el nivel de *CM*) y la detección de grupos de proteínas funcionalmente relacionadas. Para ello se compararon resultados obtenidos a diferentes niveles de *CM* (niveles 1, 5 y 10 con *p valor* $< 10^{-6}$ y $\rho' > 0{,}6$). En este punto, conviene recordar que esta aproximación parte de la idea de que la co-evolución no es necesariamente un proceso que ocurra exclusivamente entre pares de proteínas, sino que las relaciones co-evolutivas (como las funcionales) pueden darse en grupo. Por lo tanto, cabría esperar que parte de esas señales de co-evolución en grupo no fueran detectables para niveles muy específicos pero sí a niveles menos exigentes. Sin embargo, aquí surge un problema de difícil solución, ya que el nivel de *CM* ideal para cada caso dependerá del tamaño del grupo. Es por ello que se han establecido tres niveles de referencia (*CM-1*, *CM-5* y *CM-10*), considerando que cada nivel puede ser más o menos adecuado dependiendo del caso. Para explorar este punto se extrajeron subredes co-evolutivas asociadas con diferentes complejos de proteínas o rutas metabólicas conocidas. La estrategia utilizada fue recuperar todos los pares de asociaciones co-evolutivas en los que estuviera implicada al menos una proteína del complejo o ruta estudiados. De esta manera se tiene una visión de cómo se relacionan las proteínas de un complejo o ruta entre ellas (predicciones correctas), pero también con otras proteínas (predicciones incorrectas, desconocidas o pertenecientes a otro grupo).

Por ejemplo, la Figura 11A muestra las relaciones co-evolutivas detectadas para los miembros del complejo NADH deshidrogenasa tipo I. La NADH deshidrogenasa es un complejo multienzimático importante para la respiración en bacterias. Es uno de los puntos de entrada más comunes a la cadena de electrones y cataliza la transferencia electrónica desde el NADH hasta la coenzima Q. Este transporte de electrones está a su vez acoplado con la translocación de protones al inter-membrana, generando un gradiente electroquímico esencial para generación de ATP. La NADH deshidrogenasa es un complejo con forma de L compuesto en *E. coli* por 14 proteínas, que se organizan en tres subunidades estructurales y funcionales. El módulo N oxida al NADH iniciando la transferencia electrónica y está compuesto por las proteínas NuoE, NuoF y NuoG. El módulo Q, compuesto por NuoC, NuoI, NuoB y NuoD (o NuoCD), acepta electrones del módulo N y los transfiere a la coenzima Q vía centros Fe-S, siendo además el conector entre los otros dos módulos. Finalmente, el módulo P está integrado en la membrana celular, se encarga del transporte de protones a través de ella y está compuesto por NuoA, NuoH, NuoJ, NuoK, NuoL, NuoM, y NuoN (para una revisión de este sistema ver (Brandt, 2006)).

La evolución de este complejo ha sido extensivamente estudiada (Friedrich & Scheide, 2000; Mathiesen & Hägerhäll, 2003; Moparthi & Hägerhäll, 2011; Schut *et al.*, 2013).

Las 11 proteínas que componen los módulos Q y P pueden encontrarse en los tres dominios de la vida (Bacterias, Arqueas y Eucariotas), mientras que el módulo N parece ser intercambiable en función del donante de electrones. Estos estudios evolutivos sugieren que la NADH deshidrogenasa ha evolucionado de forma modular según una secuencia de pasos que comenzaría con un núcleo ancestral que incluiría al módulo Q y la mayor parte del P, al que posteriormente se añadirían NuoA, NuoJ y finalmente el módulo de N (Schut *et al.*, 2013).

Como se puede observar el la Figura 11A, éste es un caso ilustrativo de complejo que co-evoluciona como un grupo de proteínas, ya que no es posible detectar ninguna relación co-evolutiva específica (*CM-1*) para ninguna proteína del complejo. Sin embargo, a Nivel 5 (*CM-5*) es posible detectar hasta ocho señales co-evolutivas entre miembros del complejo. Es interesante que todas ellas se den entre miembros de una misma subunidad funcional y estructural (Brandt, 2006), siete en el fragmento de membrana y una en el fragmento deshidrogenasa. Aún así, cabe destacar que no se ha detectado co-evolución específica entre miembros del módulo Q. Es interesante que la separación entre el módulo P y N observada en *CM-5* es también parcialmente coherente con la evolución modular propuesta para NADH deshidrogenasa, aunque la detección de co-evolución entre NuoA y NuoJ con otros miembros del complejo P sugiere que esta segregación no se explica por la pertenencia al núcleo evolutivo o a adiciones posteriores.

Finalmente, *CM-10* predice una red altamente interconectada que contiene once miembros del complejo, pertenecientes a las tres subunidades funcionales. Esta red no contiene ninguna proteína que no pertenezca al complejo y tan solo dos proteínas del complejo no son detectadas: NuoB y NuoCD. En el caso de NuoB, es posible detectar relaciones significativas con NuoE, NuoF, NuoG, NuoH, NuoI, NuoJ y NuoK, pero que no superan el umbral de ρ' establecido (presentan valores entre 0,43 y 0,55), lo que indica que la co-evolución dentro de este módulo no es tan fuerte como en el N y el P. Por otro lado, NuoCD es un buen ejemplo de las limitaciones de nuestra estrategia para detectar ortólogos. NuoCD es una proteína que surgió por un evento de fusión génica, lo que hace que sólo se hayan detectado 18 ortólogos (en el resto de especies existen dos proteínas diferentes NuoC y NuoD), lo que muy posiblemente ha impedido detectar una señal significativa.

Otro ejemplo interesante es la maquinaria de ensamblaje flagelar. El flagelo es una máquina molecular extremadamente sofisticada que muy someramente se puede esbozar como un filamento extracelular, un cuerpo basal insertado en la membrana celular que ancla el filamento mediante una estructura en forma de codo y que contiene al motor que hace rotar al flagelo. Esta estructura ha sido objeto de un estudio evolutivo muy extenso, ya que su complejidad la ha convertido en un ejemplo de supuesta irreductibilidad evolutiva (Pallen & Matzke, 2006). Una hipótesis muy extendida es que el flagelo ha evolucionado a partir de un sistema ancestral de transporte y secreción celular (Gophna *et al.*, 2003; Pallen *et al.*, 2005). Sin embargo, un trabajo reciente sugiere que podría haber sido al revés (Abby & Rocha, 2012). Otro estudio interesante revela que el núcleo evolutivo del flagelo compuesto por 24 genes muestra una red de homologías lejanas entre 10 de ellos que podría implica que todos ellos surgieron por duplicación de un gen ancestral en el origen de bacterias (Liu & Ochman, 2007).

La maquinaria de ensamblaje flagelar en *E. coli* está compuesta por 36 proteínas según *KEGG* (Kanehisa 2004). De estas 36 proteínas *CM* es capaz de recuperar relaciones co-evolutivas para 19 de ellas, repartidas en 7 subredes, que en general representan proteínas con una relación funcional más directa (Figura 11B).

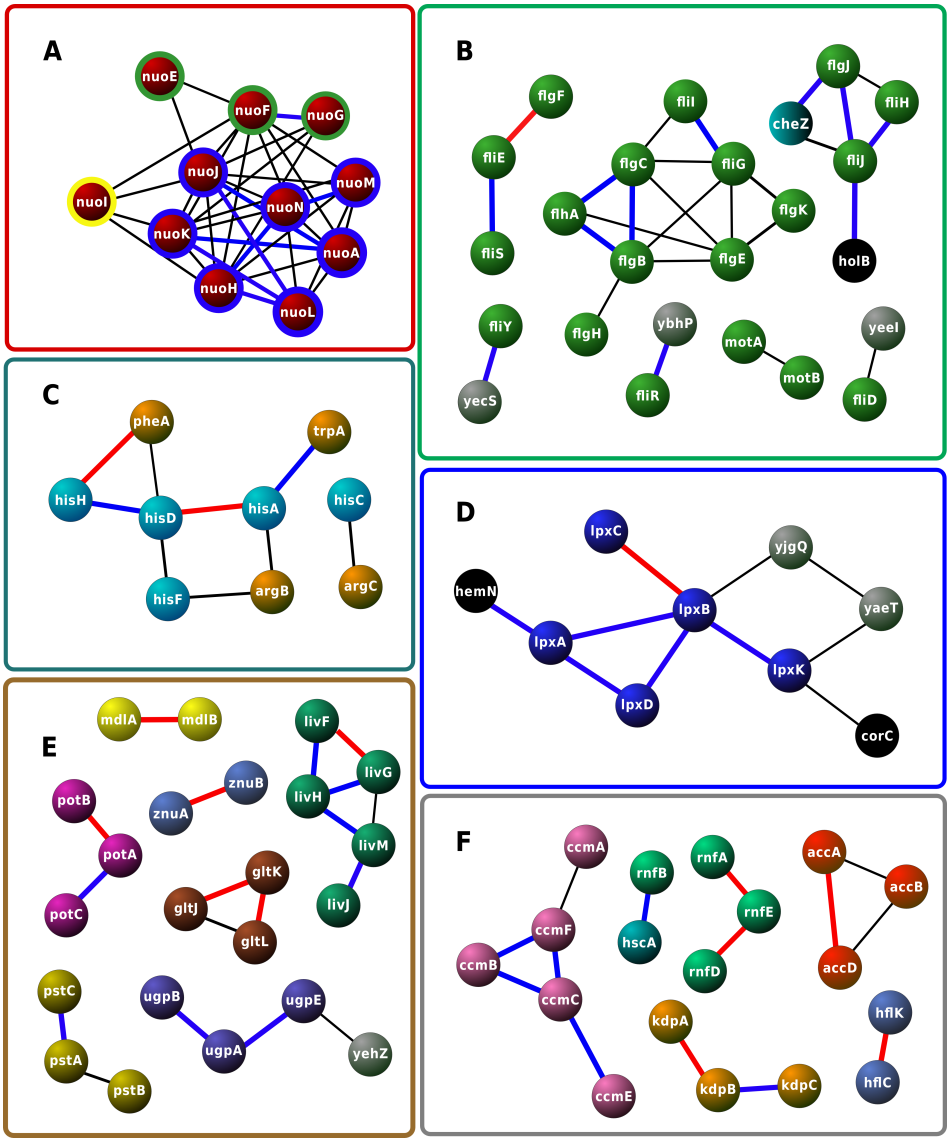


Figura 11. Ejemplos de predicciones para casos de grupos proteínas asociadas funcionalmente (*p valor* < 10^{-6} y ρ > 0,6). En cada ejemplo los nodos del mismo color corresponden al mismo grupo funcional. Los nodos coloreados en gris corresponden a nodos no caracterizados funcionalmente y los nodos negros proteínas sin relación conocida con el grupo funcional estudiado. Los enlaces entre nodos corresponden a predicciones de *CM-1* (rojo), *CM-5* (azul) y *CM-10* (negro). Los casos presentados son: **A)** Complejo NADH deshidrogenasa; **B)** Maquinaria de ensamblaje flagelar; **C)** Ruta de la biosíntesis Histidina; **D)** Ruta de la biosíntesis de Lípido A; **E)** Complejos transportadores ABC; **F)** Otros complejos moleculares. En todos los casos el grupo de proteínas relacionadas funcionalmente fue definido según *KEGG* (Kanehisa *et al.*, 2004).

En este caso *CM-1*, detecta una señal para el par de proteínas FlgF-FliE que integran el vástago de transmisión del flagelo junto con otras tres proteínas. Por otro lado, *CM-5* detecta 10 interacciones, entre las que se encuentran varios pares interesantes. Por ejemplo, FliI y FliG han sido aisladas como parte de un complejo de cinco proteínas

flagelares, aunque no parecen interaccionar directamente entre ellas (González-Pedrajo *et al.*, 2006). También se observa que FlgB, FlgC y FlhA están conectadas entre sí. De ellas FlgB y FlgC forman parte también del vástago de transmisión, mientras que FlhA se incluye en la maquinaria de secreción de las proteínas extracelulares del flagelo, incluidas FlgB y FlgC (y otras 11 proteínas más), detectándose una interacción directa entre al menos FlhB y FlgC (Minamino & Macnab, 2000).

CM-5 también agrupa a FliH, FliJ, FlgJ, CheZ y HolB a través de cuatro conexiones. Entre estas proteínas el enlace FliH-FliJ refleja una interacción física que es parte del sistema de secreción flagelar (Minamino & Macnab, 2000). La co-evolución entre FliJ y FlgJ es interesante, porque para que se dé la secreción de las proteínas flagelares, en la que participan FliH y FliJ, es necesario formar un poro en la membrana creado por FlgJ (Nambu *et al.*, 1999; Hirano *et al.*, 2001). Estos resultados sugieren que estas tres proteínas podrían haber co-evolucionado para coordinar este proceso. Las otras dos proteínas en este grupo no parecen tener una relación tan clara con estas tres proteínas. CheZ forma parte de un sistema de dos componentes quimiosensible asociado al flagelo (Parkinson *et al.*, 1983), pero no hay conexión conocida con FlgJ. Finalmente, el par FlgJ-HolB, en el que HolB es la subunidad delta de la DNA polimerasa III, parece ser una predicción errónea de relación funcional que parece provenir del poco solapamiento en las especies donde se detectan ortólogos para ellos (sólo 16 especies en común de las 116 usadas en el estudio).

Un caso particularmente interesante es la co-evolución señalada por *CM-5* entre FliY y YecS. En el momento del desarrollo de este trabajo, YecS era una proteína de función desconocida y tan sólo se pudo hallar un soporte indirecto para esta asociación (Juan *et al.*, 2008b). Sin embargo, recientemente se ha demostrado que efectivamente ambas proteínas integran un transportador ABC de L-cistina (Deutch *et al.*, 2014). Por último, *CM-5* también apunta a una posible relación funcional entre FliR y YbhP, una proteína de función desconocida.

En cuanto a *CM-10*, esencialmente recupera un grupo de 8 proteínas altamente conectadas que corresponden a diferentes elementos del flagelo y que parece constituir el núcleo de la co-evolución grupo específica de la maquinaria de ensamblaje del flagelo. Así mismo, presenta dos pares de proteínas que co-evolucionan de forma aislada. En primer lugar, detecta co-evolución entre MotA y MotB, dos proteínas que interaccionan y que forman el motor del flagelo (Kojima & Blair, 2004). Por último, *CM-10* detecta el par FliD-MtfA en el que MtfA (codificada por yeeI) es una proteína reguladora del sistema fosfotransferasa de glucosa (un sistema capaz de dirigir la quimiotaxis) (Becker *et al.*, 2006). Sin embargo, con la información disponible esta relación no parece suficiente para justificar una co-evolución entre estas proteínas. Es más, este par, al igual que el de FlgJ-HolB, parte de un solapamiento pequeño en especies (sólo 16), lo que hace que se deba ser cauto con esta predicción.

Es interesante considerar que el elevado tamaño de la maquinaria flagelar, podría requerir usar un nivel superior de *CM*. De hecho, para Nivel 20 (*CM-20*) se detectan relaciones co-evolutivas para otras cinco proteínas flagelares con la subred mayor de las detectadas a Nivel 10, sin incluir ningún nuevo falso positivo (datos no mostrados). Este ejemplo, plantea uno de los problemas en la detección de señales co-evolutivas, ya que mientras las relaciones más específicas son fácilmente distinguibles de aquellas debidas a señales inespecíficas no implicadas con interacciones funcionales, esta frontera resulta más difícil de establecer para grupos relativamente grandes de proteínas que co-evolucionan entre sí. Es decir, un nivel de *CM* demasiado alto para un grupo pequeño puede resultar adecuado para un grupo mayor.

En la Figura 11 se presentan también otros ejemplos interesantes. En primer lugar, se incluyen dos ejemplos de rutas metabólicas pequeñas: la biosíntesis de Histidina (ver Figura 11C) y la de Lípido A (ver Figura 11D). En ambos casos, se observa que las relaciones co-evolutivas son menos claras que en el caso de complejos de proteínas, mostrando una mayor tendencia a incluir falsos positivos y proteínas relacionadas indirectamente, así como a formar subredes relativamente poco interconectadas. Esto puede deberse a una combinación de un tamaño mayor de grupo funcional y a una menor co-dependencia funcional entre las proteínas de estas rutas. Finalmente, se observa una buena presencia de transportadores ABC (ver Figura 11E), probablemente debido a que representan el ideal de relaciones funcionalmente intensas y relativamente independientes del resto de las funciones celulares.

3.2. Co-evolución específica de pares de proteínas.

En esta sección se presentará una segunda metodología capaz de recuperar pares de proteínas que co-evolucionan. Esta metodología, denominada *ContextMirror Global* (*CMG*), pretende detectar aquellos pares de proteínas con una señal co-evolutiva exclusiva de ellas. Es decir, busca extraer aquellas similitudes evolutivas que no pueden ser explicadas por la *combinación de perfiles co-evolutivos* del resto de proteínas del proteoma. En este sentido, *CMG* representa un nivel mayor de exigencia que *CM-1*, que recupera aquellas similitudes evolutivas que no pueden ser explicadas por el perfil de ninguna otra única proteína.

CMG es una metodología conceptualmente similar a aquellas desarrolladas recientemente en el contexto de la co-evolución entre residuos (Weigt *et al.*, 2009; Morcos *et al.*, 2011) (ver 1. Introducción). Al igual que estas metodologías, se basa en la recuperación de la matriz global de correlaciones parciales (o su equivalente). En particular, *CMG* sigue una estrategia que consiste en corregir la matriz de partida (correlaciones en nuestro caso), para reparar las distorsiones introducidas por los problemas de muestreo y ruido, asegurando que sea invertible de forma exacta. Posteriormente se procede a dicha inversión y al cálculo de las correlaciones parciales de cada par de proteínas con respecto al conjunto de todas las demás (Cramér, 1999). *CMG* es un método computacionalmente eficiente que no requiere de grandes recursos computacionales. Esta aproximación requiere mantener la completitud e integridad de la matriz de correlaciones, para lo cual se ha desarrollado un protocolo en tres pasos, equivalente al establecido para *CM*. Este protocolo fue implementado en *R* (R Core Team, 2013) usando los paquetes *corpcor* (Schäfer *et al.*, 2015) y *Hmisc* (Harrell, 2015).

En primer lugar se desarrolló una versión de *MT* denominada *MirrorTree* estandarizado (*MTe*) que pretende recuperar la estructura global de la matriz de correlaciones entre las historias evolutivas de todas las proteínas del proteoma de referencia, controlando por los sesgos del árbol de las especies. A continuación se calcularon los *Perfiles Co-evolutivos contraídos* (*PCc*), que suponen una corrección de la matriz de correlaciones de los perfiles para compensar efectos distorsionadores de falta de cantidad u homogeneidad de los datos. Además, esta corrección asegura que la matriz sea apta para su inversión exacta. Finalmente, se procedió al paso que hemos denominado *ContextMirror Global* (*CMG*), en el que se recupera la matriz de correlaciones parciales mediante la inversión de la matriz contraída de correlaciones obtenida en el paso previo.

De forma análoga a la sección 3.1, se empezará describiendo *CMG* para acabar presentando un análisis comparativo de sus resultados y los de *CM*. En este caso, dicho

análisis estará encaminado al estudio de señales co-evolutivas más recientes recuperadas en 23 especies bacterianas diferentes.

3.2.1. *MirrorTree* estandarizado (*MTe*)

En secciones previas, se ha discutido el efecto que tiene el árbol de las especies en las similitudes entre los árboles de proteínas. Este efecto se puede resumir en dos aspectos principales: un sesgo a aumentar artificialmente los valores de *MT* y una tendencia a distorsionar algunas correlaciones como consecuencia de la aparición de distribuciones no unimodales.

Como se verá más adelante, *CMG*, a diferencia de *CM*, trata la matriz de correlaciones de *PC* en conjunto y por lo tanto es más sensible a distorsiones en su estructura. Como consecuencia, se intentó extremar el cuidado en el tratamiento de la información para reducir al mínimo estas distorsiones y evitando el uso de umbrales de confianza en los pasos intermedios.

Para ello se decidió empezar por corregir la misma matriz inicial de distancias evolutivas. Las distorsiones asociadas al sesgo filogenético provienen de situaciones en las que las distancias entre pares de ortólogos reflejan el hecho de provenir de distribuciones de distancias diferentes para cada par de especies (con diferentes medianas). Para reducir este problema se estandarizaron las distancias filogenéticas entre proteínas de dos especies dadas a lo largo de todos los árboles incluidos en el análisis. Dado que las distribuciones de estas distancias no son normales, se optó por realizar una estandarización robusta centrada en la mediana y con una escala de Desviación Absoluta de la Mediana. En aquellos casos en los que ésta es nula, se la aproximó por la Desviación Absoluta de la Media. De esta forma, se escalaron y centraron todas las distribuciones de distancias para hacerlas comparables entre ellas, reduciendo el sesgo filogenético. A continuación se calcularon las correlaciones de Pearson entre las distancias estandarizadas de cada par de proteínas, lo que proporciona la matriz de correlaciones de *MTe*. Por todo lo explicado, *MTe* puede considerarse una corrección análoga a *TMT* (Pazos *et al.*, 2005), pero que no requiere establecer un árbol de especies de referencia, que en este caso se correspondería al que se podría reconstruir usando las distancias medianas empleadas en la estandarización. Sin embargo, en este caso la corrección viene directamente extraída de los sesgos en las distancias de los árboles empleados. Además, *MTe* también escala estas distancias, lo que compensa el efecto de las diferencias en las desviaciones del valor central.

Es importante aclarar que aquí, al contrario que en *CM*, no se realiza ningún filtro por *p valor* o número mínimo de especies. Esto es así porque aunque estos valores puedan no ser totalmente comparables o correctos, resulta más importante mantener la matriz completa de correlaciones en el siguiente paso, como se explicará a continuación. Con este fin también se completaron los casos en los que no se pudo calcular la correlación (por insuficiencia de ortólogos en las mismas especies) asignándoles una correlación de cero.

3.2.2. *Perfiles co-evolutivos contraídos* (*PCc*)

Al igual que en *ContextMirror* se incorporó un paso de cálculo de correlaciones de Pearson entre perfiles co-evolutivos. En este caso, estos perfiles co-evolutivos se construyen usando la matriz completa obtenida en *MTe*. Por lo que todas las correlaciones entre perfiles se basan en el mismo número de valores (el número de proteínas en el proteoma).

Como se explicó para *CM*, las correlaciones entre perfiles co-evolutivos se usan como una aproximación de la matriz de co-dependencias evolutivas. De hecho, partiendo de la matriz completa de MTe, se obtiene una matriz de correlaciones invertible de forma exacta. Sin embargo, la eficiencia de la estimación de las correlaciones está limitada por la información de partida (el número de especies incluidas en el análisis), así como por las distorsiones heredadas de esa matriz que no se hayan podido corregir. Con el fin de mejorar esta eficiencia se utilizó una estrategia de contracción de la matriz de correlaciones (Schäfer & Strimmer, 2005; Opgen-Rhein & Strimmer, 2007). Este método realiza una contracción lineal en los valores de correlación que acerca la matriz a una matriz de referencia. De hecho la matriz contraída es una combinación de la matriz de partida (correlaciones empíricas) y la matriz de referencia (en este caso la matriz identidad). El equilibrio entre la contribución de ambas matrices lo establece la intensidad de la contracción, λ (mayores λ implican mayor contribución de la matriz de referencia). Este proceso pretende compensar el efecto de la falta de casos, que deriva en una sobrestimación de los coeficientes de correlación. Esta corrección impone que la matriz de correlaciones contraídas resultante cumpla las propiedades de ser una matriz definida positiva y bien condicionada. Estas propiedades son esenciales para poder invertir la matriz de correlaciones de forma exacta, operación necesaria para el cálculo eficiente de la matriz global de correlaciones parciales.

Uno de los puntos clave de esta aproximación es la determinación del valor de λ, ya que de él depende el nivel de corrección establecido sobre la matriz de correlaciones empíricas. En este caso se optó por una aproximación determinista en la que λ se establece analíticamente. Para ello se define el λ óptimo como aquel que minimiza el error cuadrático medio. Este λ se puede determinar como una función de la varianza de la matriz de partida, su correlación con la matriz de referencia y la diferencia cuadrática media entre ambas, según el lema establecido por Ledoit y Wolf (Ledoit & Wolf, 2003). En breve, mayores varianzas y correlaciones y menores diferencias cuadráticas producen menores valores de λ y, por lo tanto, menores contracciones en nuestra matriz. Es decir, cuando se dispone de más casos y/o la matriz se parece más a la de referencia (en este caso la matriz identidad), se contrae menos. Esta estrategia fue originalmente propuesta para mejorar la inferencia de la matriz de covarianza en situaciones con muchas variables y pocos casos (Schäfer & Strimmer, 2005). Dado que el número de especies incluidas en nuestro análisis suele ser limitado (decenas o centenares de especies), esta estrategia resulta muy útil para obtener una buena estimación de la matriz correlaciones, esencial para el cálculo global de correlaciones parciales.

3.2.3. *ContextMirror Global* (*CMG*)

El último paso de esta metodología es el cálculo de la matriz global de correlaciones parciales (en las que la dependencia de cada par de proteínas está controlada por el resto del proteoma). Para ello se obtiene la matriz de concentraciones, que es la matriz inversa de la matriz de correlaciones. La matriz de correlaciones parciales corresponde al negativo de las concentraciones (elementos fuera de la diagonal de la matriz de concentraciones) estandarizadas (Cramér, 1999).

Es importante recordar que *CMG* proporciona una estimación de la co-dependencia totalmente específica de cada par, lo cual corresponde a la señal recuperable de la co-evolución entre pares de proteínas. Por lo tanto, *CMG* es una estimación de las co-dependencias entre pares de proteínas asumiendo la ausencia de co-dependencias en grupo. En esta interpretación, las señales inespecíficas eliminadas corresponderían a la

acumulación de efectos indirectos asociados con una *red de co-dependencias real* entre pares de proteínas.

CMG contrasta con *CM*, ya que si bien representa una mayor eficiencia a la hora de recuperar co-evolución entre pares de proteínas, ésta se obtiene a costa de ignorar la información a nivel de grupos que, como se ha mostrado previamente, también podría representar una parte importante de la señal co-evolutiva.

3.2.4. Evaluación de *ContextMirror Global* y estudio de las relaciones co-evolutivas en diferentes especies

Como en el caso de *CM* se procedió a determinar la capacidad de *CMG* para recuperar pares de proteínas que interaccionan funcionalmente. De la misma forma, se utilizaron conjuntos de interacciones asociadas a co-dependencia funcional como una aproximación de pares con potencial para co-evolucionar. Como se discutió previamente, al no poder establecer un estándar de pares que realmente han co-evolucionado se compararon las diferentes aproximaciones en el contexto de sus predicciones con una señal más clara.

Tabla 1. Tabla de especies de referencia para los análisis de *CMG*

Especies	Grupo taxonómico	Número de genes	Número de especies*
Kytococcus sedentarius DSM 20547	Actinobacteria	2554	88
Brachybacterium faecium DSM 4810	Actinobacteria	3068	89
Xylanimonas cellulosilytica DSM 15894	Actinobacteria	3443	86
Beutenbergia cavernae DSM 12333	Actinobacteria	4197	74
Geodermatophilus obscurus DSM 43160	Actinobacteria	4810	77
Conexibacter woesei DSM 14684	Actinobacteria	5914	22
Capnocytophaga ochracea DSM 7271	Bacteroidetes	2171	42
Pedobacter saltans DSM 12145	Bacteroidetes	3792	35
Dyadobacter fermentans DSM 18053	Bacteroidetes	5719	89
Spirosoma linguale DSM 74	Bacteroidetes	6938	24
Chitinophaga pinensis DSM 2588	Bacteroidetes	7192	16
Bacillus subtilis subsp. subtilis str. 168	Bacilli	4176	45
Rhodobacter sphaeroides 2.4.1	Alphaproteobacteria	4242	109
Agrobacterium tumefaciens str. C58	Alphaproteobacteria	5355	81
Delftia acidovorans SPH-1	Betaproteobacteria	6040	117
Stigmatella aurantiaca DW4/3-1	Deltaproteobacteria	8352	7
Helicobacter pylori 26695	Epsilonproteobacteria	1594	24
Acinetobacter sp. ADP1	Gammaproteobacteria	3307	189
Vibrio cholerae O1 biovar El Tor str. N16961	Gammaproteobacteria	3834	169
Escherichia coli str. K-12 substr. MG1655	Gammaproteobacteria	4146	183
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2	Gammaproteobacteria	4525	180
Pseudomonas putida KT2440	Gammaproteobacteria	5350	189
Pseudomonas aeruginosa PAO1	Gammaproteobacteria	5571	195

En este caso la evaluación de *CMG* se desarrolló como un análisis de la co-evolución en 23 proteomas bacterianos de 7 filos diferentes. Estas especies fueron elegidas por el proyecto europeo *MICROME* (http://www.microme.eu/) por su interés biotecnológico, biomédico o medioambiental. Por lo tanto, estas 23 especies constituyen un excelente conjunto para evaluar la robustez y aplicabilidad de nuestros métodos en diferentes especies bacterianas (ver Tabla 1). En este contexto resulta particularmente interesante, estudiar la señal co-evolutiva detectable para un conjunto de especies filogenéticamente más relacionadas con cada especie de referencia.

En trabajos realizados en colaboración con el grupo del Dr. Florencio Pazos (Herman *et al.*, 2011), se había observado que la elección del grupo de especies puede afectar a la detección de señales co-evolutivas. En particular este trabajo mostró que los pares de proteínas entre los que se detectaba co-evolución dependían de la divergencia entre las especies incluidas. De hecho, esta selección de especies determina la antigüedad y extensión de la señal detectada. Es decir, enfocarse en divergencias muy grandes (como en la evaluación de *CM*) implica que las relaciones detectadas deben ser antiguas y conservadas en gran parte de la evolución de procariotas. Por otro lado, utilizar especies evolutivamente próximas recuperará pares de proteínas que han co-evolucionado más recientemente en ese grupo de especies. Así mismo, dicho trabajo evidenció la necesidad de equilibrar la divergencia entre las distintas especies incluidas, para lo que se estableció un protocolo de selección de especies relativamente próximas a la de referencia (ver Materiales y Métodos).[†]

Es importante tener presente que en el marco propuesto se han realizado 23 análisis independientes, que arrojaron información acerca del comportamiento de los métodos desarrollados en diferentes contextos evolutivos. Sin embargo, el nivel de conocimiento experimentalmente soportado disponible para muchas de estas especies es en general bajo. Por lo tanto, se decidió comparar la capacidad predictiva de *CMG* con *CM* a dos niveles diferentes. En primer lugar se observó el comportamiento de nuestros métodos en *E. coli* utilizando algunos de los conjuntos de referencia empleados previamente para evaluar *CM* (complejos bien establecidos, rutas metabólicas e interacciones físicas a pequeña escala; ver Materiales y Métodos). Posteriormente se realizó una comparación a lo largo de las 23 especies utilizando conjuntos de relaciones funcionales establecidos por la base de datos *STRING* (Szklarczyk *et al.*, 2015). Estos conjuntos son una combinación de información experimental e inferencias, por lo que se decidió establecer criterios muy exigentes de confianza (ver Materiales y Métodos).

3.2.4.A. Comparación de CMG y CM en la evolución reciente de Escherichia coli

La comparación de la capacidad predictiva de la aplicación de nuestros métodos a un conjunto de 183 Gammaproteobacterias evolutivamente próximas a *E. coli* arroja unos resultados interesantes (ver Figura 12). En primer lugar, a pesar de la menor divergencia evolutiva, los resultados para *CM-1* y *CM-10* son comparables a los obtenidos previamente con especies mucho más divergentes. Esto muestra que un número considerable de relaciones funcionales han dado lugar a co-evolución entre proteínas también más recientemente. Igual de interesante es el bajo nivel de correlación entre los resultados de *CM-1* y *CM-10* en éste análisis y en el realizado sobre especies más divergentes (r =0,09 y *p valor* < 2,2 x 10^{-16} en ambos casos). Estos

[†] Por comparación con el análisis presentado en la sección 3.1, denominaremos los análisis realizados con especies del mismo grupo taxonómico como de co-evolución reciente, aunque el término reciente aquí puede implicar tiempos bastante largos (ver Materiales y Métodos).

resultados muestran que, aunque en ambos casos se recuperan pares informativos, estos pares son muy diferentes. De hecho, de las 970 predicciones de *CM-10* con $r>0.6$ y *p valor*$<10^{-6}$ en el análisis anterior, sólo 43 son recuperadas entre las 2.453 predicciones de *CM-10* (con el mismo criterio) en este análisis. Estos resultados son compatibles con los obtenidos en una exploración previa de la importancia de la selección de especies (Herman *et al.*, 2011) y confirma que análisis diseñados a diferentes escalas evolutivas pueden proporcionar información complementaria.

Sin embargo, lo que resulta más llamativo es la mejora predictiva que presenta *CMG* para todos los tipos de relaciones funcionales estudiados (ver Figura 12). Por ejemplo, las primeras 2.000 predicciones de *CMG* presentan una precisión superior a 0,9 para proteínas en el mismo complejo, frente a menos de 0,8 en *CM-1* y *CM-10*. Situación que se confirma a nivel de rutas metabólicas e incluso en interacciones físicas donde se pasa de una precisión de alrededor de 0,2 para *CM-1* y *CM-10* a casi de 0,6 para *CMG* en las primeras 1000 predicciones.

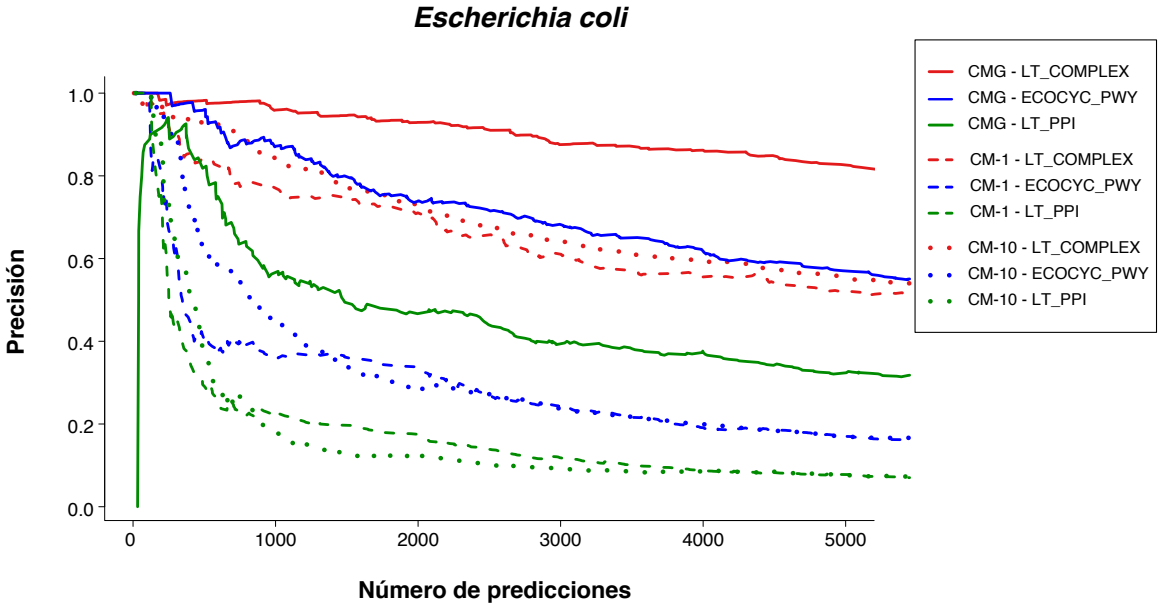


Figura 12. Precisión de *CM-1*, *CM-10* y *CMG* para *E. coli* empleando especies de su grupo taxonómico. La precisión es evaluada mediante los conjuntos de complejos bien establecidos (*LT_COMPLEX*), rutas metbólicas (*ECOCYC_PWY*) e interacciones físicas determinadas por experimentos a pequeña escala (*LT_PPI*).

Es particularmente ilustrativo el comportamiento de los métodos respecto al conjunto de interacciones físicas, ya que estas interacciones es probable que estén asociadas a casos de co-evolución específica de par. Por lo tanto, la mejora observada de *CMG* frente a *CM* en este conjunto es una muestra de la mayor eficiencia de *CMG* para recuperar la señal co-evolutiva a nivel de pares de proteínas.

Es también llamativo que esta mejora en la detección de co-evolución entre pares de proteínas se traduzca a su vez en una mejoría para los conjuntos de referencia asociados a grupos de proteínas funcionalmente inter-dependientes (*i.e.* complejos de proteínas y rutas metabólicas). Este efecto muestra que *CM-1* además de perder señales de grupos de proteínas no recupera parte de la señal específica de pares entre los miembros de estos grupos que *CMG* sí es capaz de rescatar.

Aún más interesante es que, a pesar de los buenos resultados de *CMG* es conjuntos de grupos funcionales, las predicciones obtenidas por *CMG* y *CM-10* para el mismo sistema (evolución reciente de *E. coli*) son también poco solapantes, ya que sólo 406 predicciones de *CM-10* están entre las primeras 5.000 predicciones de *CMG*. Esta situación es compatible con la idea de que *CM-10* recupera asociaciones en grupo que no se reflejan en los métodos más específicos como *CM-1* y *CMG*.

Estos resultados muestran la superioridad predictiva de *CMG* para la evolución reciente de *E. coli* e ilustran la influencia de la definición del contexto evolutivo y la complementariedad de ambas aproximaciones.

3.3.4.B. Comparación de CMG y CM en la evolución reciente de 23 especies diferentes

A continuación se comparó la capacidad predictiva en las 23 especies bacterianas seleccionadas por *MICROME* (http://www.microme.eu/). Para ello se evalúo la precisión de *CMG*, *CM-1* y *CM-10* para cada una de las 23 especies, utilizando como conjunto de referencia las interacciones funcionales establecidas como muy fiables (*puntuación*>=900, ver Materiales y Métodos) por *STRING* (Szklarczyk *et al.*, 2015). Las interacciones funcionales recogidas en *STRING* representan un tipo de relación que combina interacciones físicas, rutas metabólicas, co-regulación y otras relaciones inferidas por varios métodos a lo largo de muchas especies diferentes. Por lo tanto, este conjunto se consideró como el más apropiado para evaluar la precisión de nuestros métodos en un grupo tan variado de especies.

Este análisis ofrece dos resultados interesantes. En primer lugar confirma que *CMG* presenta una mayor capacidad predictiva que *CM-1* y *CM-10* en todo el conjunto analizado de condiciones de número de especies incluidas y grupos taxonómicos (ver Figura 13, Figura 14 y Figura 15). De hecho esta mejora es robusta y consistente a lo largo de sistemas de especies muy diferentes y empleando distintos conjuntos de especies de grupos taxonómicos muy diferentes.

Así mismo, muestra que el número de especies incluidas en el análisis es clave para mejorar las predicciones. De hecho, cuatro de las cinco especies con peores predicciones para *CMG* (Figura 15, Tabla 1) incluyeron menos de 25 especies en sus análisis (la restante incluyó 44 especies), mientras seis de las siete especies con mejores predicciones incluyeron más de 100 especies (la restante incluyó 81 especies).

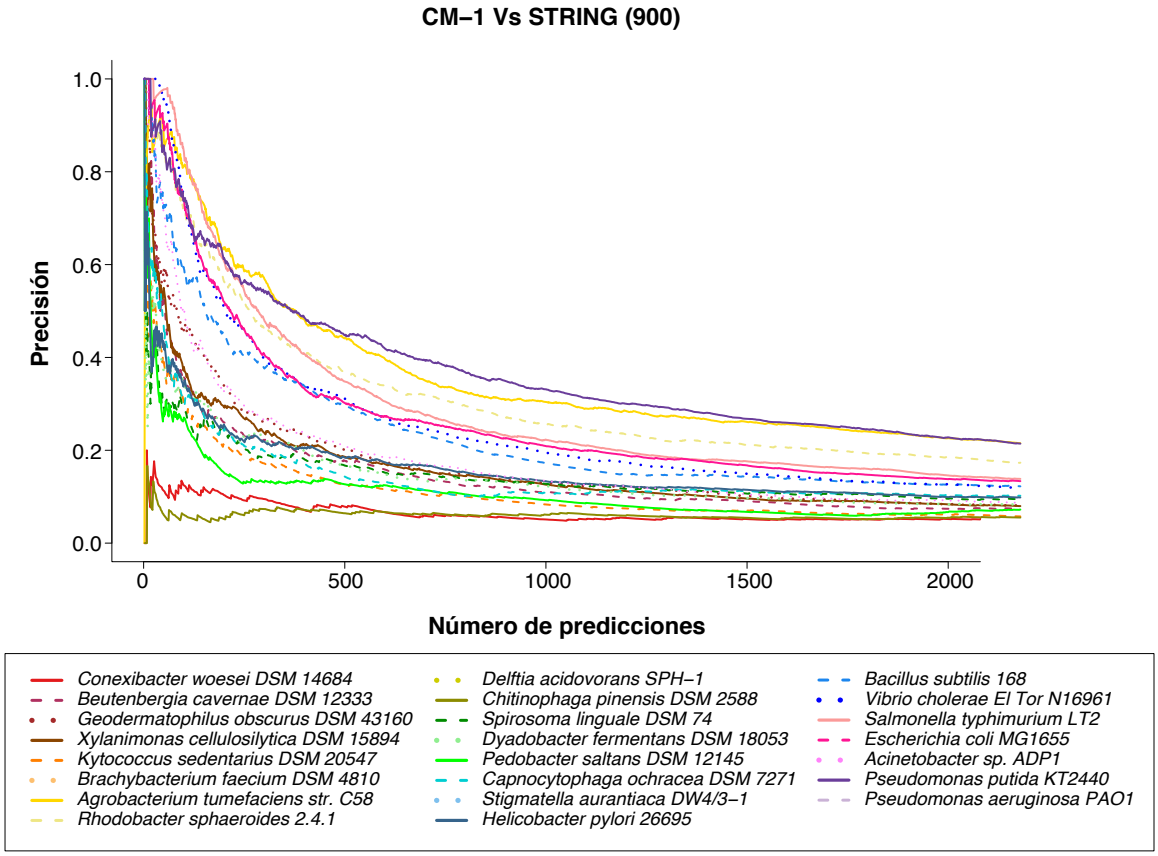


Figura 13. Precisión de *CM-1* en 23 especies bacterianas. La precisión es evaluada frente a la interacciones funcionales establecidas por *STRING* (Szklarczyk *et al.*, 2015) con una confianza mayor de o igual a 900 (con un máximo de 999). No se representan las interacciones de *Stigmatela aurantica* por no disponer de sufientes especies (se requieren 16 especies) para realizar predicciones (ver Tabla 1).

La influencia del número de especies incluidas en el análisis condiciona claramente los resultados por grupos taxonómicos, obteniéndose las mejores predicciones de *CMG* para Gamma, Beta y Alphaproteobacterias, con niveles intermedios para Actinobacterias, Bacteroidetes y Firmicutes y los niveles más bajos para Epsilon y Deltaproteobacterias. Sin embargo, en todos los casos, *CMG* recupera señales co-evolutivas informativas con mayor eficiencia y robustez que *CM-1* y *CM-10*. Tan sólo en el caso extremo *S. aurantiaca*, para la que sólo se pudieron recuperar 7 especies, *CMG* no fue capaz de recuperar una señal co-evolutiva asociada a interacciones funcionales.

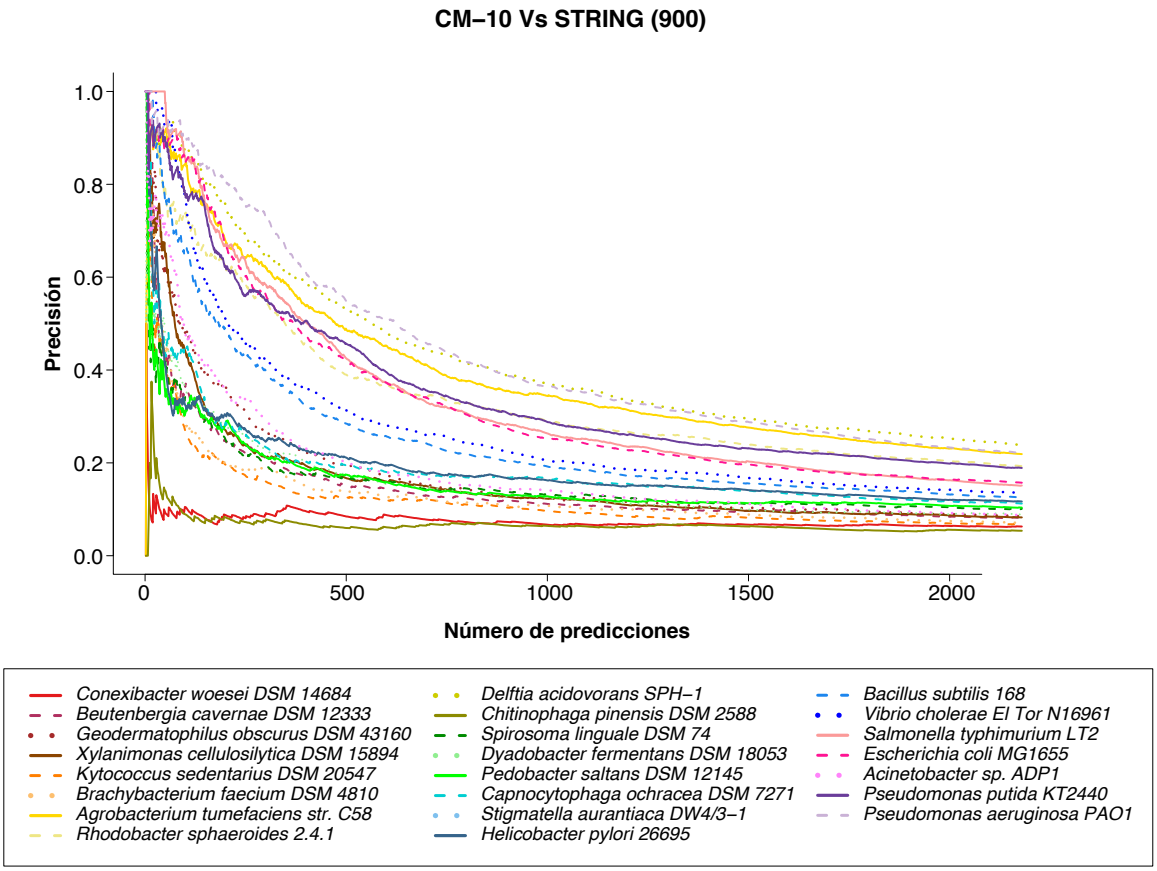


Figura 14. Precisión de *CM-10* en 23 especies bacterianas. La precisión es evaluada frente a la interacciones funcionales establecidas por *STRING* (Szklarczyk *et al.*, 2015) con una confianza mayor de o igual a 900 (con un máximo de 999). No se representan las interacciones de *Stigmatela aurantica* por no disponer de suficientes especies (se requieren 16 especies) para realizar predicciones (ver Tabla 1).

En este punto, es importante señalar que, como muestra la Tabla 1, la relación entre el número de especies (menos de 200) y el de proteínas (más de 2000) es extremadamente pequeña para este tipo de métodos basados en la obtención de la matriz de correlaciones parciales (o equivalente). Baste recordar que predicciones de alta calidad a nivel de residuos requieren un número de secuencias no redundantes comparable al de posiciones (de 500 a 1000 secuencias) (Morcos *et al.*, 2011).

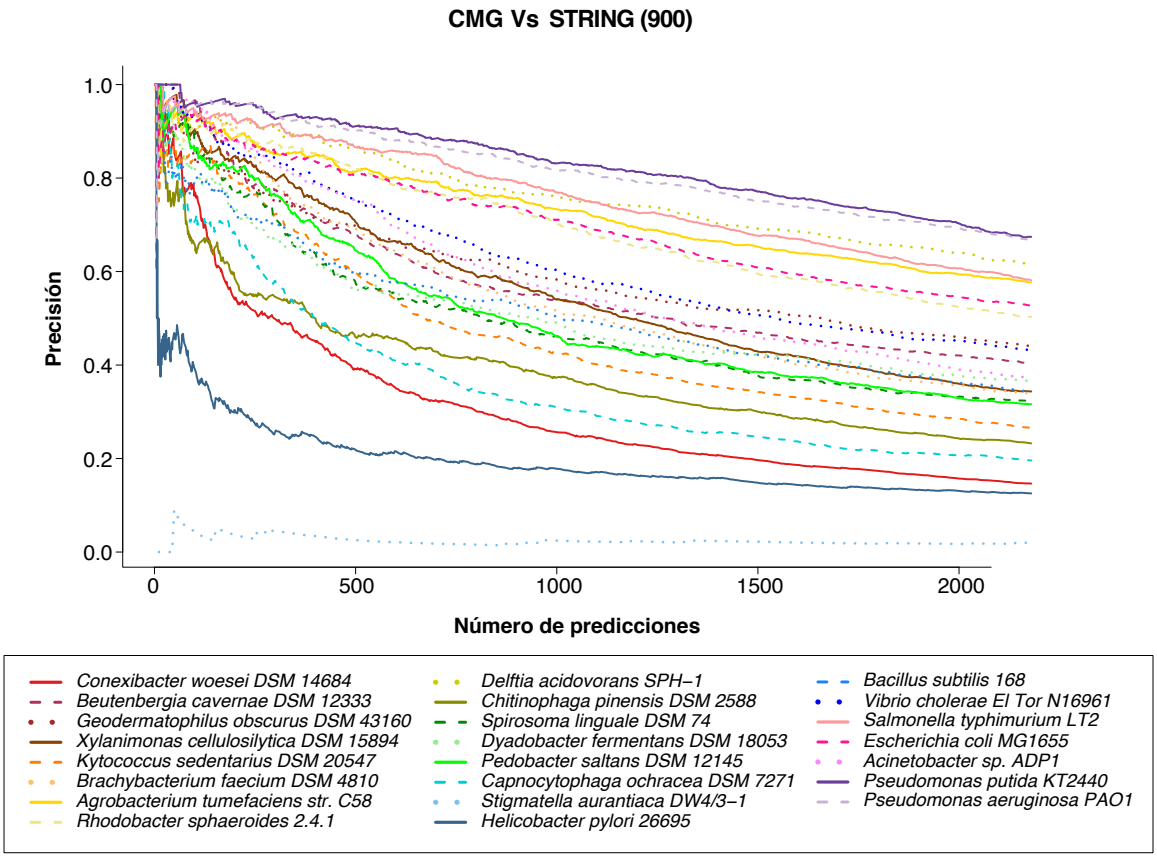


Figura 15. Precisión de *CMG* en 23 especies bacterianas. La precisión es evaluada frente a la interacciones funcionales establecidas por *STRING* (Szklarczyk *et al.*, 2015) con una confianza mayor de o igual a 900 (con un máximo de 999). Aunque se representan las interacciones de *Stigmatela aurantica*, los pobres resulados obtenidos para ella se explican por no disponer de sufientes especies para realizar predicciones fiables (ver Tabla 1).

3.2.4.C. Análisis de los términos GO sobrerrepresentados en proteínas detectadas por ContextMirror Global en las 23 especies

Como se ha comentado anteriormente, las predicciones obtenidas en *E. coli* incluyendo especies distantes o próximas son muy diferentes. Por lo tanto, resulta interesante estudiar en qué medida los pares detectados en la evolución reciente de especies diferentes apuntan a los mismos procesos celulares y funciones moleculares. Para ello, se establecieron las primeras 500 predicciones confirmadas por *STRING* como los conjuntos adecuados para comparar predicciones de diferentes especies. Este criterio permite estudiar qué relaciones funcionales presentan señales más claras de co-evolución reciente, eliminando el efecto de los falsos positivos cuya contribución variaría en función de las condiciones de cada análisis.

En primer lugar se realizó un análisis de enriquecimiento de términos de ontología génica (términos *GO*, (Ashburner *et al.*, 2000; Szklarczyk *et al.*, 2015)). Este análisis determina qué anotaciones génicas están sobre-representadas en un conjunto de genes en función de lo esperado según la distribución de términos asociados a todos los genes de cada especie (obtenidos de *UniProt-GOA*; (Barrell *et al.*, 2009; Szklarczyk *et al.*, 2015)). Para ello se utilizó *GO-TermFinder* (Boyle *et al.*, 2004), con el que se obtuvo la

lista de términos *GO* sobre-representados (*p valor* < 0.01 y *FDR* < 1%) en las proteínas implicadas en las interacciones funcionales detectadas por *CMG* para cada uno de las 23 especies (ver Métodos).

Como resultado se obtuvieron un gran número de términos *GO* de Función Molecular (85 términos), Proceso Biológico (159 términos) y Componente Celular (48 términos). Como se puede observar en las Figura 16A, Figura 17A y Figura 18A, las distribuciones del número de especies en que aparece un término sigue un patrón de muchos términos *GO* únicos y muy pocos comunes (hasta en un máximo de 12 especies).

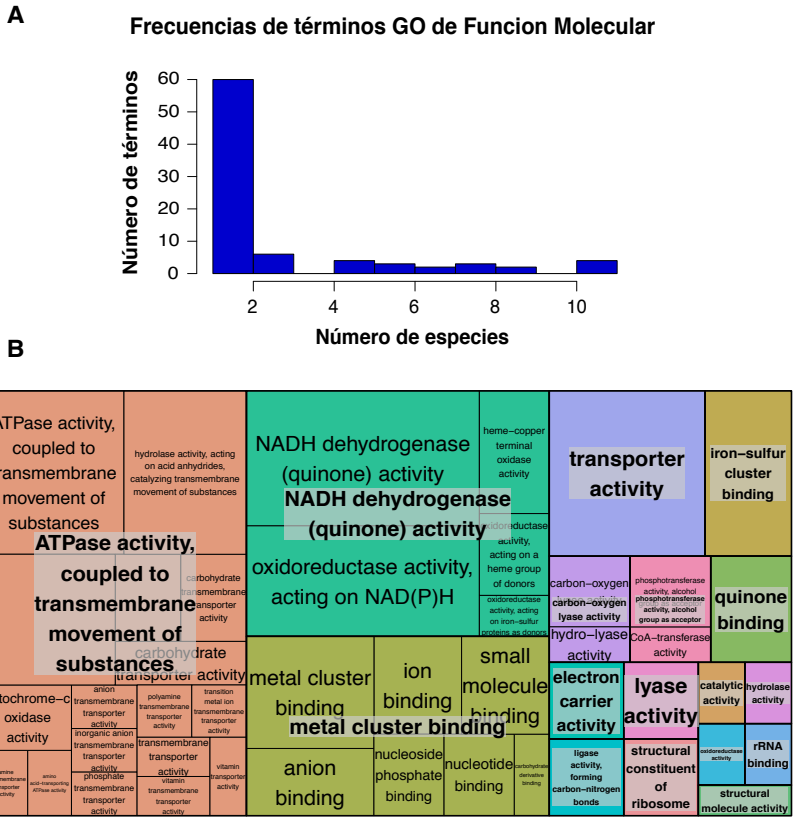


Figura 16. Términos *GO* de Función molecular enriquecidos en predicciones de *CMG* para 23 especies bacterianas. Se presentan conjuntamente los análisis de sobre-representación de términos *GO* de Función Molecular realizados con *GO TermFinder* (Boyle *et al.*, 2004) (*p valor* < 0.01 y *FDR* < 1%) A) Histograma del número de términos sobre-representados en diferentes números de especies. Se observa que la mayoría de los términos son específicos de especie. B) Representación que resume las frecuencias mostradas en A) mediante una agrupación de las anotaciones en una jerarquía de dos niveles de especificidad de los términos *GO*. En esta representación cada término *GO* mostrado ocupa un rectángulo de área proporcional a la frecuencia de sobre-representación en distintas especies. Así recuadros mayores implican una mayor frecuencia de anotaciones asociadas a la mostrada. Los colores representan el nivel de anotación más general y los términos agrupados por esa anotación se muestran como rectángulos del mismo color que se reparten el área manteniendo la proporcionalidad descrita. Las funciones moleculares más frecuentes son: actividad ATPasa acoplada al movimiento de sustancias a través de la membrana; actividad NADH deshidrogenasa; centro de unión a metales y actividad transportadora. Esta representación fue obtenida utilizando *REVIGO* (Supek *et al.*, 2011).

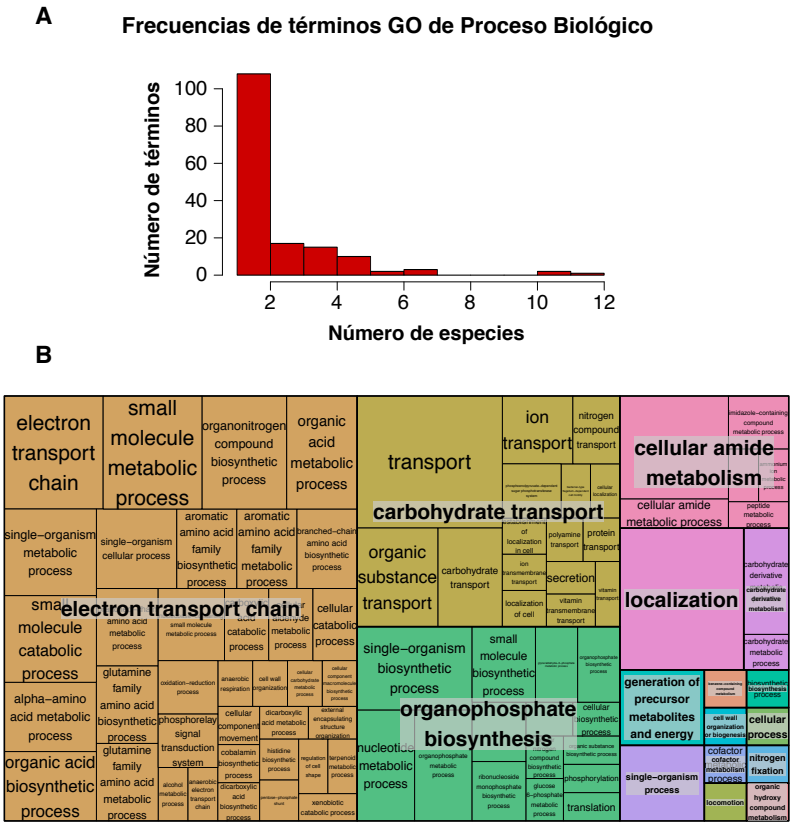


Figura 17. Términos *GO* de Proceso biológico enriquecidos en predicciones de *CMG* para 23 especies bacterianas. Se presentan conjuntamente los análisis de sobre-representación de términos *GO* de Proceso Biológico realizados con *GO TermFinder* (Boyle *et al.*, 2004) (*p valor* < 0.01 y *FDR* < 1%) A) Histograma del número de términos sobre-representados en diferentes números de especies. Se observa que la mayoría de los términos son específicos de especie. B) Representación que resume las frecuencias mostradas en A) mediante una agrupación de las anotaciones en una jerarquía de dos niveles de especificidad de los términos *GO*. En esta representación cada término *GO* mostrado ocupa un rectángulo de área proporcional a la frecuencia de sobre-representación en distintas especies. Así recuadros mayores implican una mayor frecuencia de anotaciones asociadas a la mostrada. Los colores representan el nivel de anotación más general y los términos agrupados por esa anotación se muestran como rectángulos del mismo color que se reparten el área manteniendo la proporcionalidad descrita. Los procesos biológicos más frecuentes son: cadena de transporte de electrones; transporte de carbohidratos; biosíntesis de organofosfatos y metabolismo de amidas. Esta representación fue obtenida utilizando *REVIGO* (Supek *et al.*, 2011).

Sin embargo, estos términos no son independientes sino que forman parte de una estructura ontológica que les relaciona. Con la intención de obtener un buena visión global de los términos recuperados a lo largo de las diferentes especies se optó por utilizar *REVIGO* (Boyle *et al.*, 2004; Supek *et al.*, 2011). *REVIGO* resume los términos de *GO* utilizando una medida de similitud semántica que considera la estructura de la ontología de referencia (en nuestro caso *simRel* (Schlicker *et al.*, 2006)) y establece los más representativos de cada grupo en función del valor asociado al término por el

usuario y de su especificidad. En nuestro caso, se empleó el número de especies en las que cada término apareció sobre-representado.

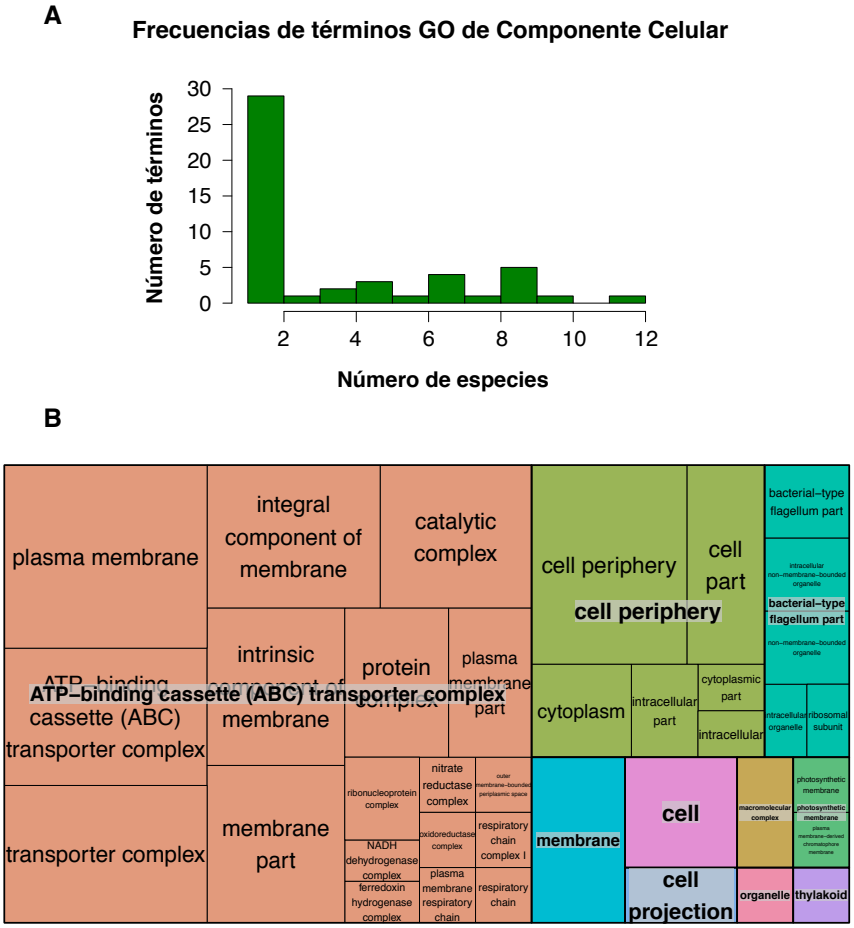


Figura 18. Términos *GO* de Componente celular enriquecidos en predicciones de *CMG* para 23 especies bacterianas. Se presentan conjuntamente los análisis de sobre-representación de términos *GO* de Componente Celular realizados con *GO TermFinder* (Boyle *et al.*, 2004) (*p valor* < 0.01 y *FDR* < 1%) A) Histograma del número de términos sobre-representados en diferentes números de especies. Se observa que la mayoría de los términos son específicos de especie. B) Representación que resume las frecuencias mostradas en A) mediante una agrupación de las anotaciones en una jerarquía de dos niveles de especificidad de los términos *GO*. En esta representación cada término *GO* mostrado ocupa un rectángulo de área proporcional a la frecuencia de sobre-representación en distintas especies. Así recuadros mayores implican una mayor frecuencia de anotaciones asociadas a la mostrada. Los colores representan el nivel de anotación más general y los términos agrupados por esa anotación se muestran como rectángulos del mismo color que se reparten el área manteniendo la proporcionalidad descrita. Los componentes celulares más frecuentes son: complejo transportador ABC, periferia celular y flagelo de tipo bacteriano. Esta representación fue obtenida utilizando *REVIGO* (Supek *et al.*, 2011).

En las Figura 16B, Figura 17B Figura 18B se muestran representaciones "*Treemap*" (Supek *et al.*, 2011) de los resultados de *REVIGO* para cada tipo de términos *GO*. En

esta representación, los términos obtenidos por *REVIGO* se agrupan en uno más general, obteniendo una jerarquía de dos niveles de términos. En nuestro caso, esta jerarquía se representa repartiendo la superficie de un rectángulo entre los términos generales y la de estos entre los más específicos de forma proporcional al número de especies en las que se detectaron como sobre-representados. En este gráfico resulta evidente que hay determinados procesos, funciones y componentes que son mucho más comunes. De hecho, el transporte de electrones y de carbohidratos, así como los complejos (transportadores ABC) y funciones asociados a ellos (actividades ATPasa, NADH deshidrogenasa, transportadora, acarreadora de electrones, etc.) son muy comunes. Así mismo, también son comunes algunas funciones metabólicas como la biogénesis de organofosfatos o el metabolismo de amidas. Igualmente interesante es la presencia del flagelo como componente sobre-representado, que también está fuertemente asociado a procesos de transporte y quimiotaxis.

Esta visión de los resultados parece sugerir que a pesar de la variedad de términos *GO*, existen unos procesos que están habitualmente afectados por co-evolución reciente a lo largo de muchas especies, tales como los procesos de transporte y obtención de energía. Es fácil imaginar que estos procesos van a ser muy importantes para definir la capacidad de adaptación de las diferentes especies. Sin embargo, estos análisis no dejan claro en qué medida los pares concretos que co-evolucionan en las diferentes especies corresponden a los mismos grupos de ortólogos en las diferentes especies, o si afecta a diferentes proteínas más allá de la diferente composición de los distintos proteomas.

3.2.4.D. Comparación de las redes co-evolutivas en las diferentes especies

Para poder comparar las diferentes redes se recuperaron las relaciones de ortología de *eggNOG* (Powell *et al.*, 2014). *eggNOG* proporciona una definición de ortología más inclusiva que la utilizada por *CM* y *CMG*, y por lo tanto más apta para detectar similitudes entre las redes co-evolutivas de dos o más especies. Utilizando estas relaciones de ortología se estableció el nivel de solapamiento en los pares co-evolucionando según el criterio utilizado en el análisis de *GO*. De esta manera definió la tasa de solapamiento como el cociente del número de pares ortólogos co-evolucionando en ambas especies, frente al número de pares co-evolucionando (500 pares en este caso).

En la Figura 19 se observa que las tasas de solapamiento son muy pequeñas (menores a 0,2) para la inmensa mayoría de los casos. Aún así, estas tasas muestran que existe una coherencia entre estas tasas de solapamiento y la divergencia evolutiva entre las especies, ya que un agrupamiento jerárquico de estas tasas agrupa las especies según su grupo taxonómico (con la excepción de aquellos grupos que sólo contienen una especie). A pesar de que esta señal es bastante débil (el agrupamiento contiene ramas internas muy cortas) se observan algunos grupos claros de especies con un solapamiento relativamente alto (entre 0,25 y 0,41). Entre ellas está un grupo de tres Actinobacterias (*X. cellulosilytica*, *B. faecium* y *B. cavernae*) y un grupo de cuatro Bacteroidetes (*S. linguale*, *D. fermentans*, *P. saltans* y *C. pinensis*), así como los pares: *S. enterica-E. coli* (Gammaproteobacteria); *R. Sphaeroides-A. tumefaciens* (Alphaproteobacteria).

Estos resultados muestran que si bien parece haber determinados procesos en los que la co-evolución ha contribuido recientemente en muchas especies, los pares concretos que sufren esta co-evolución son muy diferentes entre especies. Así mismo, como era de esperar, estas diferencias también reflejan la divergencia evolutiva entre las especies analizadas. Con la intención de entender mejor esta situación se estudió más en detalle

la co-evolución en tres procesos diferentes: la fosforilación oxidativa, los sistemas de transporte y el ensamblaje flagelar. Estos procesos, además de encontrarse entre los procesos y estructuras sobre-representados en el análisis de términos *GO*, suponen una ampliación de los casos también estudiados en detalle usando *CM*.

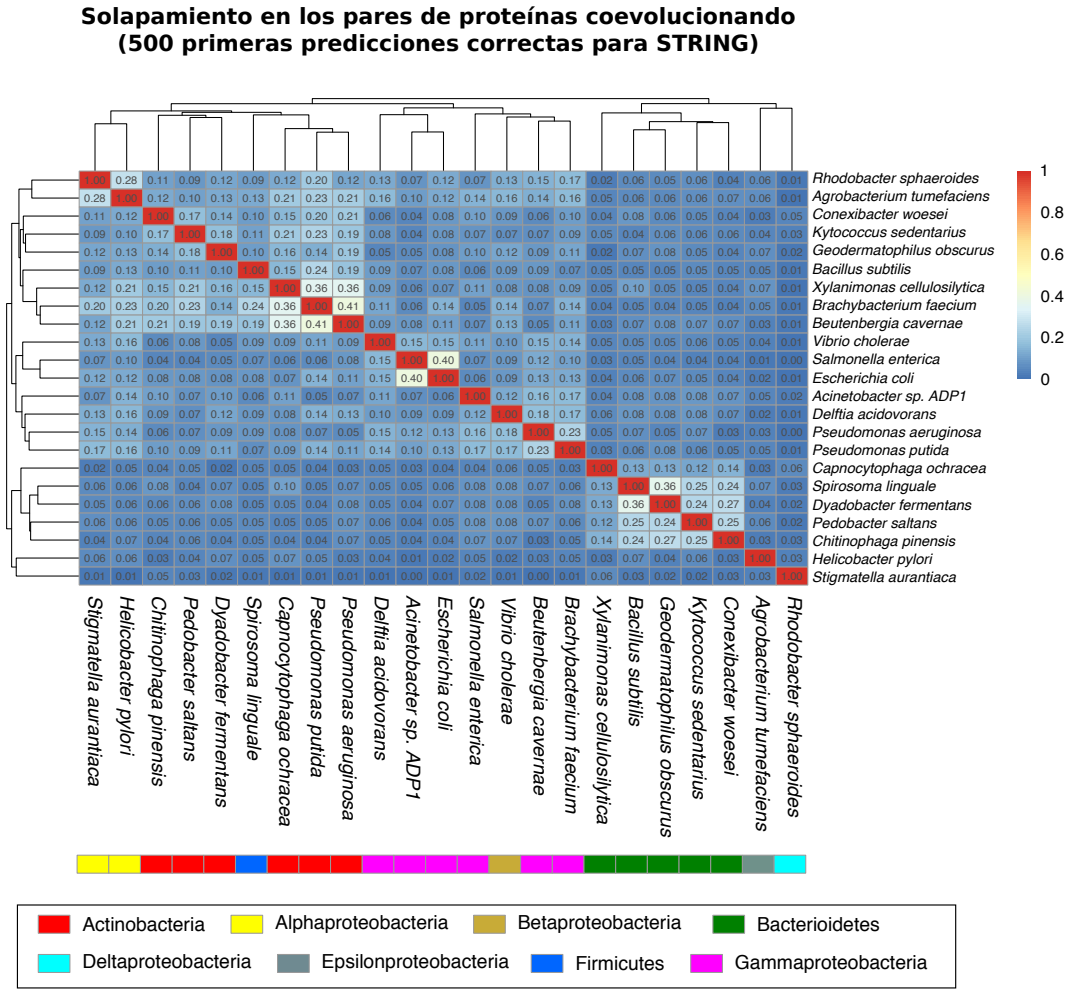


Figura 19. Solapamiento entre las predicciones de *CMG* para 23 especies bacterianas. Clasificación doble (*biclustering*) de la matriz de las tasas de solapamiento detectadas entre diferentes especies de las primeras 500 predicciones que apuntan a interacciones funcionales según *STRING* (Szklarczyk *et al.*, 2015). El gradiente de color de las celdas de la matriz es proporcional al valor de la tasa de solapamiento. El grupo taxonómico de cada especie es indicado con el código de colores representados en la parte inferior de la figura. Se observan unas tasas de solapamiento bajas pero capaces de agrupar las especies pertenecientes a los grupos taxonómicos más poblados. La figura fue obtenida utilizando las opciones por defecto (distancias euclídeas y método de vínculo completo) de la función *pheatmap* del paquete de *R* (R Core Team, 2013) *pheatmap* (Kolde, 2015).

3.2.4.E. Estudio de la co-evolución detectada por ContextMirror Global para las diferentes especies en tres procesos concretos

Para estudiar estos sistemas se obtuvieron las definiciones de las proteínas implicada en estos procesos de la base de datos *KEGG* (Kanehisa *et al.*, 2004) (ver Materiales y Métodos). *KEGG* contiene asignaciones a estos procesos para 22 de las 23 especies analizadas (no contiene *Pedobacter saltans*), por lo que se estudió la co-evolución reciente de estos procesos en esas 22 especies. En este caso, dado que interesa comparar los resultados para sistemas bastante complejos a lo largo de muchas especies, se utilizaron las señales de co-evolución detectadas por *CMG* entre los miembros de estos sistemas confirmadas por *STRING*.

La fosforilación oxidativa es un proceso que aparece señalado por el análisis de *GO* con la sobrerrepresentación de términos asociados a la NADH(P) deshidrogenasa o el transporte de electrones. Por lo tanto, se decidió analizar el proceso completo (ver Figura 20). El transporte de electrones está formado por una cadena de reacciones de oxidación y reducción que conlleva el transporte de electrones entre diferentes grupos donantes y aceptores. Así mismo, esta cadena está acoplada a fenómenos de transporte de iones (típicamente H^+) a través de la membrana celular que dan lugar a un gradiente electroquímico posteriormente aprovechado en la síntesis de ATP. Por razones de sencillez se organizaron los complejos implicados en la cadena de transporte según la disposición de la cadena mitocondrial. Así, se describe la NADH deshidrogenasa como la integrante del grupo I (homóloga del complejo I mitocondrial), la succinato y la fumarato deshidrogenasas como grupo II, las citocromo deshidrogenasas como grupo III, las citocromo oxidasas como grupo IV y las ATP sintasas como grupo V. Según esta organización, las diferentes cadenas de electrones en bacterias podrían involucrar a diferentes combinaciones de complejos, pero se puede considerar como la secuencia típica la que comienza en los grupos I y/o II, continua en el III, para terminar en el grupo IV. Finalmente el grupo V, que no forma parte de la cadena electrónica, sintetiza ATP a partir del gradiente electroquímico generado por la cadena. Este es un esquema general orientativo, pero en bacterias existe una gran diversidad de combinaciones de complejos y cadenas diferentes a ésta.

En primer lugar, llama la atención que en todas las especies estudiadas se detectan señales co-evolutivas entre pares de proteínas directamente involucradas en la fosforilación oxidativa. Entre ellas, la especie con mayor número de pares co-evolucionando es *Helicobacter pylori* con 51 pares repartidos entre la NADH deshidrogenasa (41), la citocromo c oxidasa tipo cbb3 (6), el complejo citocromo c-ubiquinol reductasa (3) y la fumarato reductasa (1). En el otro extremo, *Conexibacter woesei* tan sólo muestra co-evolución en dos pares de la NADH deshidrogenasa.

También es interesante el elevado grado de especificidad de las señales co-evolutivas confirmadas, ya que tan sólo 13 de las 337 señales de co-evolución detectadas entre proteínas que participan en la fosforilación oxidativa involucran a proteínas de complejos diferentes (ver Figura 20B). De estas 13 señales inter-complejo, 9 implican co-evolución entre proteínas los complejos del grupo III y del grupo IV (contiguos en la cadena de transporte de electrones). Otra de estas señales inter-complejo conecta la NADH deshidrogenasa con la citocromo reductasa (grupo III), que también son contiguas en la cadena.

Las tres últimas co-evoluciones inter-complejo corresponden a la co-evolución de CoxA y CoxB, miembros de la citocromo C oxidasa, con CyoE. CyoE está anotada en *KEGG* como parte de la citocromo o ubiquinol oxidasa, sin embargo, aunque forma

parte del mismo operón, CyoE está implicada en la síntesis del grupo hemo (Saiki *et al.*, 1993b) y no formar parte de este complejo (Saiki *et al.*, 1993a; Boyle *et al.*, 2004), aunque podría ser un factor en la biogénesis de la citocromo c oxidasa. De hecho, CyoE tiene una señal más débil de co-evolución con el resto de proteínas asignadas al complejo que éstas entre sí. CyoE parece co-evolucionar más comúnmente con la subunidad I de las citocromo oxidasas (CyoB y CoxA). Esta subunidad contiene los grupos hemo, que en el caso de la CyoB son los hemo B y O, cuya transformación cataliza CyoE (Saiki *et al.*, 1993b). En el caso de CoxA el grupo hemo es el A, que se sintetiza a partir del O, por lo que esta co-evolución podría estar asociada al equilibrio de concentración entre grupos hemo en especies que contienen ambas citocromo oxidasas. En esta dirección es interesante que en *B. subtilis* y *R. Sphaeroides*, el ortólogo de CyoE, interaccione con el de CtaA (la sintasa del grupo hemo A) (Brown *et al.*, 2004), la cual ha co-evolucionado con CoxC.

Como era de esperar, la NADH deshidrogenasa muestra una gran acumulación de señales de co-evolución en muchas de las especies estudiadas (hasta 15 especies diferentes de las 17 donde está presente), convirtiéndola en un complejo importante en la co-evolución reciente de muchas especies. De hecho, más de la mitad de las señales de co-evolución en este sistema (186 de 337 pares) se dan entre miembros de la NADH deshidrogenasa (ver Figura 20B). Es interesante que la distribución de estas co-evoluciones muestra una mayor concentración en los módulos N y P, de forma semejante a lo observado por *CM*. El módulo Q, para el que no se pudo recuperar señales con *CM*, aparece también con menos señales intra-módulo que inter-módulo. Esto refuerza la idea de que la co-evolución entre los miembros del módulo Q es menos común que entre los módulos N y P y sugiere que su evolución está muy condicionada por su papel de conector de los otros módulos.

Otros complejos que han co-evolucionado recientemente en muchos casos son la succinato deshidrogenasa (14 especies de 21) o la F_0F_1-ATP sintasa (12 especies de 22). De hecho en estos complejos la co-evolución implica a diferentes pares de proteínas en cada especie. Por ejemplo en la F_0F_1-ATP sintasa, el par ATPF1D/ATPF1G sólo ha co-evolucionado en *Stigmatella aurantiaca* (una Deltaproteobacteria) mientras ATPF0B/ATPF1E sólo lo ha hecho en *Kytococcus sedentarius* (una Actinobacteria) y ATPF0A/ATPF1G sólo en *Capnocytophaga ochracea* (un Bacteroidetes). Por otro lado, el par ATPF0A/ATPF0C ha co-evolucionado en ocho especies diferentes de tres grupos taxonómicos diferentes (Actinobacteria, Bacteroidetes y Gammaproteobacteria).

En el resto de los complejos implicados en la fosforilación oxidativa se observó que están involucrados en procesos de co-evolución reciente más o menos específicos de especie. Es decir, que sólo algunas especies recuperan señales de co-evolución entre ellas. Esto abarca un rango muy amplio de situaciones. Por ejemplo, la citocromo b ubiquinol oxidasa ha co-evolucionado en 11 especies (en todos los grupos taxonómicos con más una especie analizada). En el otro extremo aparecen la citocromo aa3-600 menaquinol oxidasa y la menaquinol-citocromo c reductasa que sólo han co-evolucionado en *Bacillus subtilis*, ya que es la única especie de las analizadas que contiene estos sistemas (según *KEGG*).

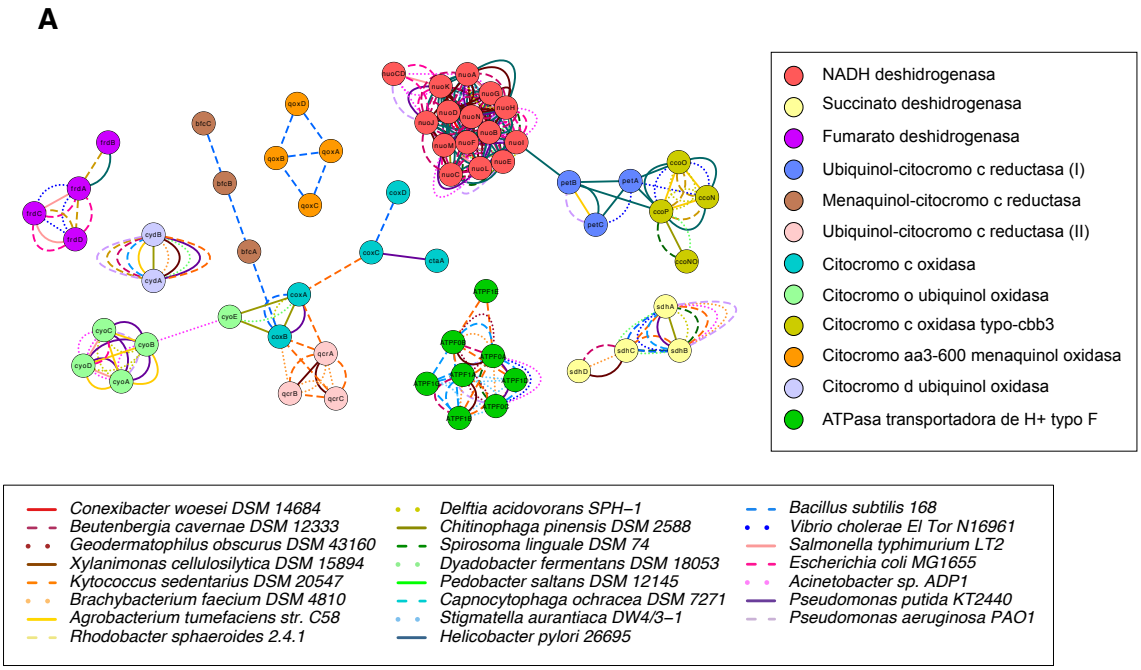


Figura 20. Predicciones de *CMG* para los complejos de la fosforilación oxidativa en 23 especies bacterianas. Las predicciones mostradas corresponden a las primeras 500 predicciones de *CMG* que apuntan a interacciones funcionales según *STRING* (Szklarczyk *et al.*, 2015). **A**) Red de las predicciones de *CMG*. Los colores de los nodos representan diferentes complejos involucrados en la fosforilación oxidativa según *KEGG* (Kanehisa *et al.*, 2004). La combinación del tipo de línea y color de los enlaces indican la especie donde se detectó la asociación. Aparecen dos complejos citodromo c oxidasas correspondientes a dos grupos de ortólogos diferentes. Se observa poco solapamiento de las predicciones de *CMG* para diferentes especies, y pocas interacciones entre complejos diferentes. La figura se obtuvo utilizando *Cytoscape* (Shannon *et al.*, 2003). **B)** (página siguiente) Matriz de solapamientos de las predicciones en diferentes especies. El gradiente de color de las celdas es proporcional al número de especies donde se detecta la interacción. Se observa una mayor presencia de interacciones detectadas en más especies en posiciones cercanas a la diagonal, lo que refleja señales de co-evolución más comunes dentro del mismo módulo funcional de la NADH deshidrogenasa (excepto para el módulo Q), dentro del mismo complejo o en complejos que realizan pasos consecutivos en la cadena de transporte de electrones (predicciones entre complejos I y III y entre complejos III y IV).

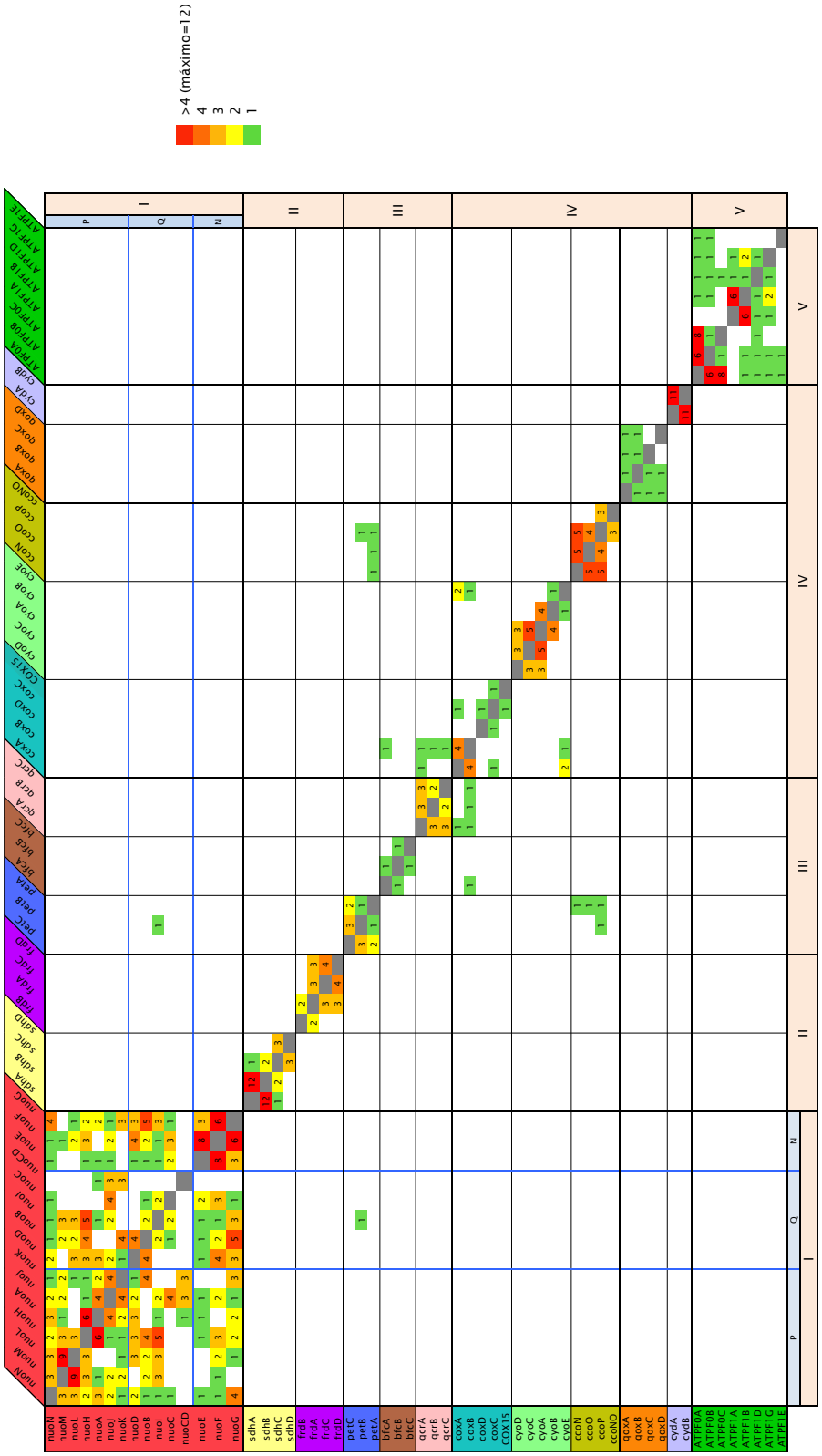


Figura 20. Predicciones de *CMG* para los complejos de la Oxidación Fosforilativa en 23 especies bacterianas. (cont.)

Otro tipo de procesos claramente señalado por el análisis de *GO* es el transporte a través de la membrana celular. Estos procesos también son clave para la adaptación de las bacterias al medio mediante el intercambio de compuestos útiles o perjudiciales para la célula. De hecho se recuperan señales co-evolutivas de hasta 88 sistemas de transporte diferentes incluidos in *KEGG* (ver Materiales y Métodos). En la Figura 21 se muestran los 52 complejos con mayor número de relaciones co-evolutivas diferentes.

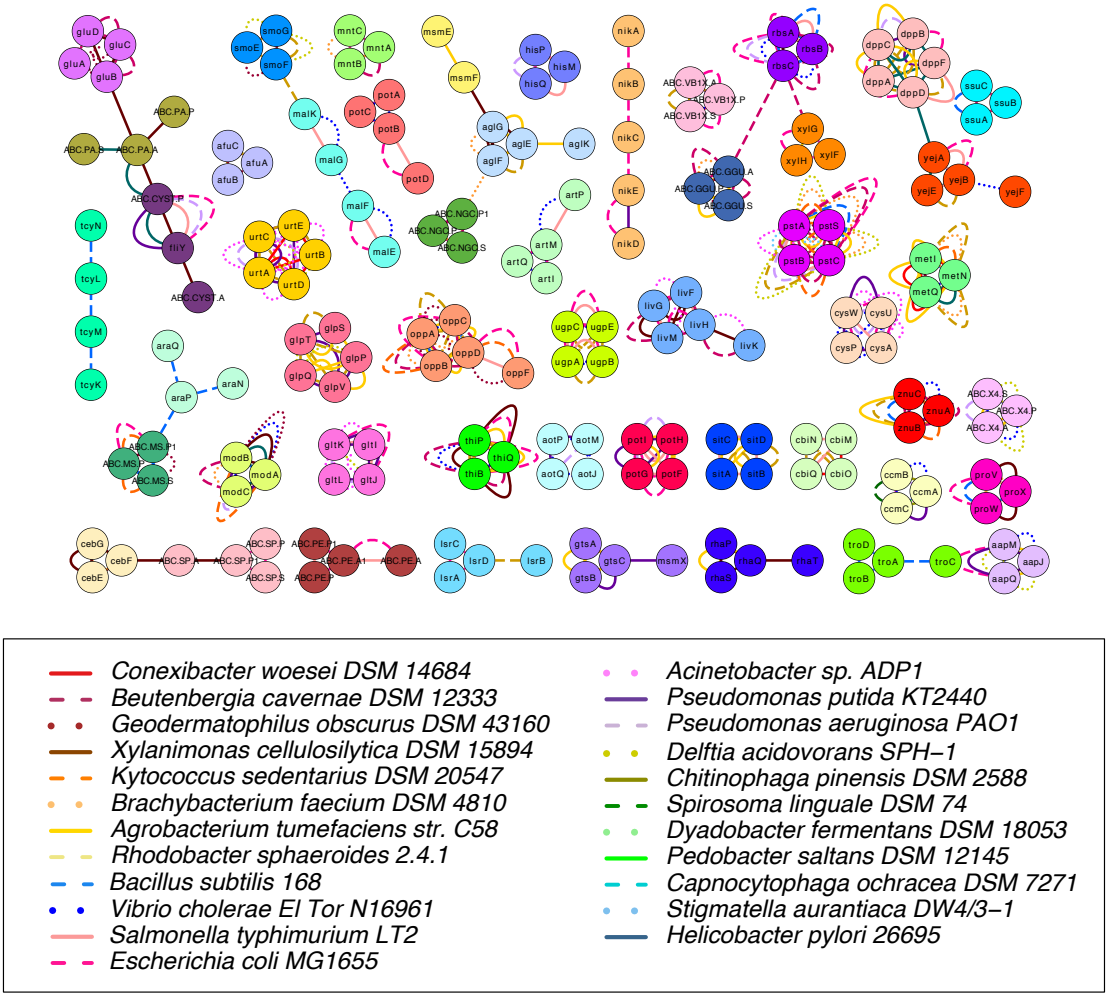


Figura 21. Predicciones de *CMG* para complejos transportadores de membrana en 23 especies bacterianas. Se representa la red de las predicciones de *CMG* para los 52 transportadores de membrana según *KEGG* (Kanehisa *et al.*, 2004) con mayor número de predicciones (3 o más). Las predicciones mostradas corresponden a las primeras 500 predicciones de *CMG* que apuntan a interacciones funcionales según *STRING* (Szklarczyk *et al.*, 2015). Los colores de los nodos representan diferentes complejos transportadores de membrana. La combinación del tipo de línea y color de los enlaces indican la especie donde se detectó la asociación. Se observa poco solapamiento de las predicciones de *CMG* para diferentes especies, y pocas interacciones entre complejos diferentes. La figura se obtuvo utilizando *Cytoscape* (Shannon *et al.*, 2003).

Al igual que en el caso de la Fosforilación Oxidativa, se encuentran algunos sistemas de transporte cuya co-evolución reciente está muy extendida en diferentes especies. Entre

ellos están el sistema de transporte de fosfatos (14 especies de 6 grupos taxonómicos diferentes), el de transporte de oligopéptidos (10 especies de 4 grupos), el de transporte de molibdato (10 especies de 4 grupos) y el de transporte de D-metionina (9 especies de 5 grupos). Sin embargo, la mayoría de los sistemas de transporte sólo co-evolucionan en un grupo pequeño de especies como el sistema de transporte de lactosa/L-arabinosa, el de L-cisteína y el de manganeso/zinc/hierro (los tres sólo en *Bacillus subtilis*), el de hierro (III) (en *Geodermatophilus obscurus*), el de Autoinductor 2 (en *Escherichia coli* y *Rhodobacter Sphaeroides*) o el de Ramnosa (en *A. tumefaciens* y *X. cellulosilytica*).

Los sistemas de transporte ilustran de forma muy clara la situación planteada por el análisis de términos *GO*. Es decir, existen muchos casos de co-evolución entre sus proteínas (hasta 610 en las 22 especies), pero existe mucha diversidad en qué sistemas y pares han co-evolucionado recientemente en cada una de las especies.

Finalmente se estudió cómo ha co-evolucionado el ensamblaje flagelar en las distintas especies. El ensamblaje flagelar es un proceso que, aunque aparece sobre-representado en los términos *GO* (ver Figura 18), sólo está presente en 12 de las 22 especies presentes en *KEGG*. Esto sugiere que debe ser un proceso importante en estas especies. De hecho, 10 de estas 12 especies muestran co-evolución en él. Entre ellas llama la atención *G. obscurus* con 50 pares de proteínas co-evolucionando. Estos pares involucran hasta 20 de las 24 proteínas asociadas al ensamblaje flagelar en esta especie. Otro caso similar es el de *H. pylori* con 43 asociaciones co-evolutivas entre 22 de las 28 proteínas implicadas en el ensamblaje flagelar en ella. Curiosamente, a pesar de estos números, hay un solapamiento relativamente bajo entre los pares recuperados para estas dos especies (sólo 12 pares). Esto vuelve a recalcar la diversidad de pares sobre los que puede actuar la co-evolución en distintos contextos evolutivos. De hecho, sólo dos pares de proteínas, MotA-MotB (forman el motor) y FlhA-FlhB (dos proteínas esenciales del sistema de secreción del flagelo) co-evoluciona en un máximo de cinco especies diferentes.

Esta situación podría sugerir que hay poca relación entre los pares para los que se detecta co-evolución y el nivel de relación funcional entre ellos. Es decir, que se estarían obteniendo pares de la maquinaria del ensamblaje flagelar sin una relación funcional o estructural particularmente fuerte. Sin embargo, como se puede observar en la Figura 22B, las proteínas implicadas en papeles o estructuras locales (según KEGG) están más conectadas entre ellas y en más especies. De hecho, de los nueve grupos de proteínas que reflejan subestructuras del ensamblaje del flagelo en siete todos los miembros de la subestructura co-evolucionan en al menos una especie. La primera de las otras dos subestructuras es el filamento con su capuchón, en el que el FliC (que forma el filamento) muestra pocas señales de co-evolución con otras proteínas estructuralmente próximas en el flagelo. La segunda subestructura corresponde a la que involucra a más proteínas, el sistema de secreción tipo III. En este sistema se detectan hasta 11 pares de proteínas co-evolucionando en alguna especie y las 7 proteínas de este sistema co-evolucionan con al menos otra del grupo. Esta observación es similar a la obtenida en los casos analizados en detalle utilizando *ContextMirror* (ver sección 3.1) y sugiere que la estructura subyacente de co-dependencias funcionales y espaciales condiciona la probabilidad de detectar señales de co-evolución. Para la comprensión de estos resultados y de las diferencias de la co-evolución en especies distintas es importante tener en mente la variedad estructural que presenta la estructura flagelar en diferentes especies (Chen *et al.*, 2011).

Las dos proteínas involucradas en más pares que co-evolucionan (con 15 de las 29 que lo componen) son FliM (proteína del anillo C que actúa como conmutador del sentido

de rotación del motor) y FlhA (del sistema de secreción flagelar). Además, FlhA es la proteína que co-evoluciona en más ocasiones con proteínas flagelares a lo largo de las 22 especies (25 pares). De hecho, aunque la mayoría de las señales de co-evolución se explican por proximidad en el flagelo (103 de las 158 co-evoluciones detectadas en el ensamblaje del flagelo; ver Figura 22B), este sistema de secreción muestra un elevado número de señales de co-evolución no explicados por proximidad en la estructura del flagelo (31 de las 55 en esa situación). En este sentido, es interesante que se haya determinado experimentalmente que FlhA (y en muchos casos también FlhB) interacciona con los miembros del sistema de secreción FlhB, FliI, FliH, FliJ, FliO, FliP y FliQ (Minamino & Macnab, 2000; McMurry *et al.*, 2004), con el anillo MS (FliF) (Kihara *et al.*, 2001) y con FlgB, FlgD, FlgE, FlgK, FlgL, FliC (Minamino & Macnab, 2000), con la chaperona FlgN (Minamino *et al.*, 2012) y con los complejos FliD/FliT y FliC/FliS (Kanehisa *et al.*, 2004; Bange *et al.*, 2010; Kinoshita *et al.*, 2013). Este elevado número de interacciones muestra la importancia del sistema de secreción y explica la variedad de otras subestructuras con las que co-evoluciona.

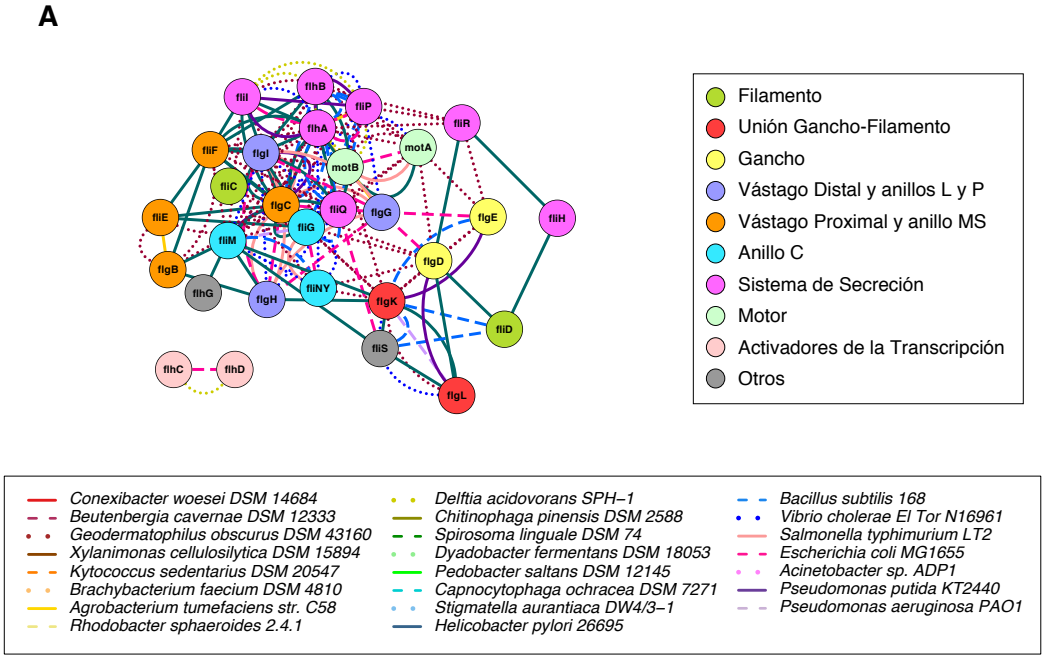


Figura 22. Predicciones de *CMG* para proteínas del *Ensamblaje Flagelar* en 23 especies bacterianas. Las predicciones mostradas corresponden a las primeras 500 predicciones de *CMG* que apuntan a interacciones funcionales según *STRING* (Szklarczyk *et al.*, 2015). **A)** Red de las predicciones de *CMG*. Los colores de los nodos representan diferentes módulo funcionales o estructurales involucrados en el ensamblaje flagelar según *KEGG* (Kanehisa *et al.*, 2004). La combinación de tipo de línea y color de los enlaces indican la especie donde se detectó la asociación. Se observa poco solapamiento de las predicciones de *CMG* para diferentes especies, y pocas interacciones entre complejos diferentes. La figura se obtuvo utilizando *Cytoscape* (Shannon *et al.*, 2003). **B)** Matriz de solapamientos se las predicciones en diferentes especies. El gradiente de color de las celdas es proporcional al número de especies donde se detecta la interacción. Se observa una mayor presencia de interacciones detectadas en más especies en posiciones cercanas a la diagonal, lo que refleja señales de co-evolución más comunes dentro del mismo módulo funcional o en módulos que se encuentran próximos estructuralmente según el modelo canónico de KEGG (área delimitada por la línea roja).

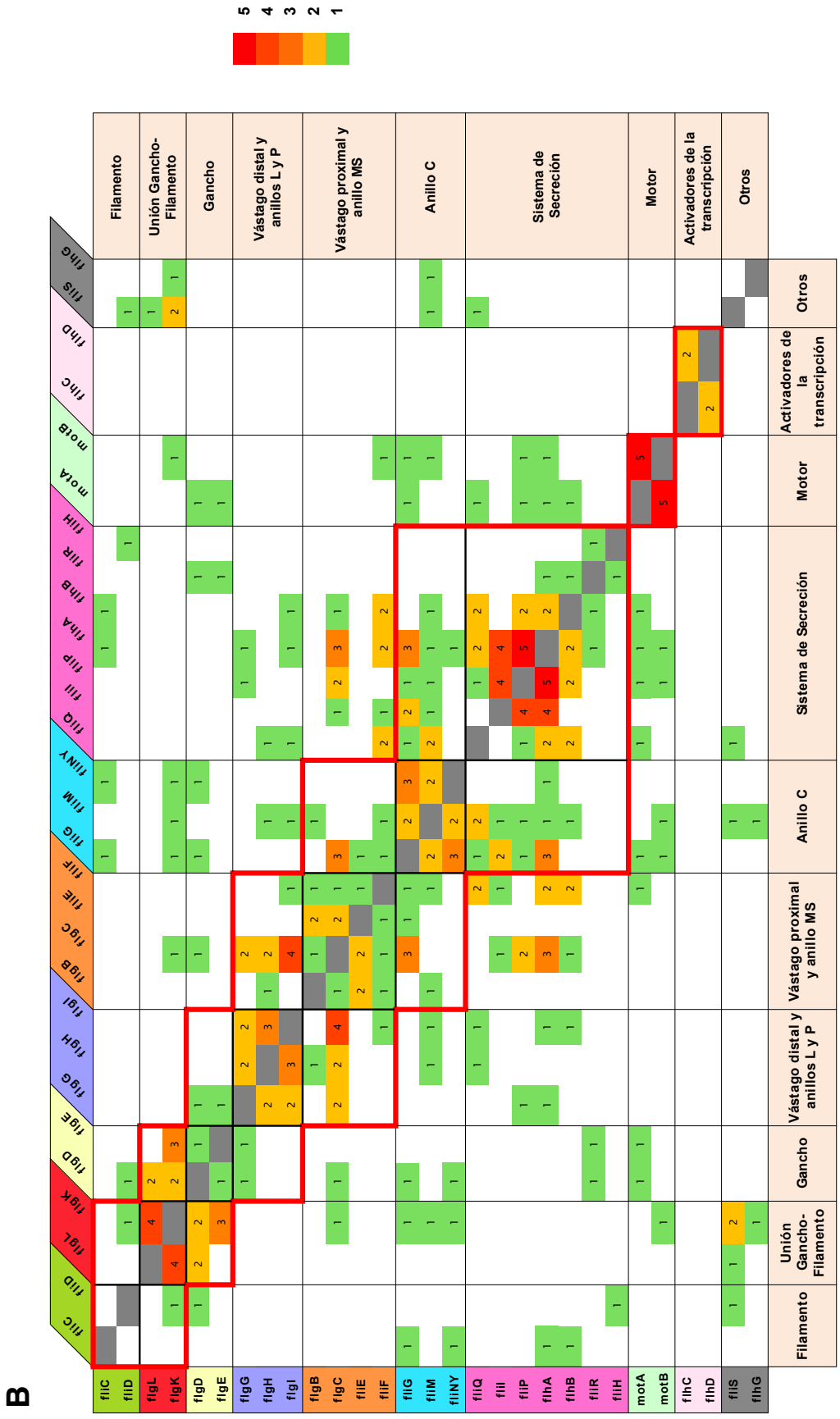


Figura 22 . Predicciones de *CMG* para proteínas del Ensamblaje Flagelar en 23 especies bacterianas. (cont.)

4. Discusión

En esta memoria se han presentado dos metodologías para la detección de señales co-evolutivas entre proteínas predictivas de interacciones: *ContextMirror* (*CM*) y *ContextMirror Global* (*CMG*). Estos métodos son herederos directos de los trabajos previos en los que se proponía el uso de las similitudes entre los árboles de proteínas para la predicción de interacciones entre ellas (método *MirrorTree* o *MT*).

La principal aportación conceptual de *CM* y *CMG* consiste en que ambos consideran el proteoma completo como el objeto de análisis adecuado para discernir qué señales de parecido en los árboles de proteínas están asociadas a co-evolución. De esta forma, el sistema analizado pasa de ser el par de proteínas a ser el proteoma completo. Este cambio de paradigma nos permite plantear métodos que consideran las señales de similitud entre árboles como una composición de efectos propios y otros ajenos a cada par de proteínas, relacionados con la influencia de otras proteínas del resto del proteoma. Este marco conceptual se desarrolla mediante el análisis de correlaciones parciales entre los árboles de todas las proteínas de una especie de referencia. Desde un punto de vista metodológico, los dos métodos descritos en este trabajo, *CM* y *CMG*, implementan dos tratamientos muy distintos de las contribuciones propias y ajenas a la co-evolución de un par de proteínas.

4.1. *ContextMirror* detecta señales de co-evolución compartidas entre grupos de proteínas

CM está diseñado como un método para explorar diferentes grados de especificidad de la señal de similitud en la historia evolutiva de cada par de proteínas. Es decir, para detectar señales compartidas con terceras proteínas, que puedan estar asociadas a fenómenos de co-evolución en grupo. La evaluación de los resultados obtenidos por *CM* en *E. coli* demuestra que esta aproximación permite detectar señales co-evolutivas indicativas de co-evolución entre los miembros de un mismo complejo o ruta metabólica de forma mucho más precisa. Esta mejora se observa en todos los conjuntos evaluados de interacciones físicas y funcionales, aunque es particularmente evidente en el caso de complejos extraídos manualmente de la literatura. Así mismo, los resultados muestran que admitir la influencia de hasta diez compañeros co-evolutivos permite recuperar señales no detectadas por niveles más específicos. Esto sugiere la existencia de fenómenos de co-evolución en grupo. Finalmente, la combinación de *CM* con datos de aislamiento de complejos a gran escala ha permitido recuperar un subconjunto de interacciones, soportadas por ambas metodologías, con mayor nivel de confirmación respecto a anotaciones manuales que la de los datos experimentales de partida.

Además el estudio de casos concretos muestra que diferentes niveles de especificidad de *CM* pueden reflejar diferentes niveles de relación estructural dentro de un complejo. Así, en la NADH deshidrogenasa se observa que los niveles más específicos no pueden detectar ninguna asociación. Admitir cinco compañeros co-evolutivos revela asociaciones dentro de los módulos funcionales y estructurales del complejo. Finalmente, admitir diez compañeros permite recuperar un elevado número de asociaciones dentro del complejo, sin incluir ninguna predicción fuera del él.

4.2. *ContextMirror Global* detecta señales co-evolutivas específicas del par de proteínas

CMG ha sido diseñado para extraer con mayor eficiencia la señal co-evolutiva asociada a similitudes entre árboles filogenéticos totalmente exclusiva del par de proteínas. Esta aproximación, inspirada en métodos recientemente desarrollados para la predicción de contactos (Weigt *et al.*, 2009; Morcos *et al.*, 2011) es aún más precisa que *CM* en todos los conjuntos evaluados de interacciones funcionales en *E. coli*. En particular, llama la atención su mayor capacidad para recuperar interacciones físicas entre proteínas, más difíciles de detectar por otros métodos basados en *MT*. Además, la mayor precisión de *CMG* en la predicción de relaciones funcionales de grupos de proteínas, sugeriría que también en éstas existen señales específicas de pares. En principio, parece difícil determinar si la señal detectada como específica por *CMG* proviene de la descomposición en pares de la señal del grupo. Sin embargo, la baja coincidencia en los pares predichos por *CM* y *CMG* para *E. coli.* refuerza la hipótesis de que ambos métodos detectan señales diferentes y complementarias. Cualquiera que sea el caso, es importante señalar la importante mejora en las predicciones obtenidas por *CMG*.

También se han presentado los resultados de *CMG* en 23 especies bacterianas diferentes enfocándose en la co-evolución detectable en sus respectivos grupos taxonómicos. La evaluación de estos resultados confirma la mayor precisión de *CMG* en la predicción de interacciones funcionales en muy diferentes condiciones. Estos resultados muestran una dependencia de la cantidad de especies empleadas en el análisis, aunque con buenos resultados para todos los casos en los que se emplearon más de 25 especies.

Una comparación de los casos de interacciones funcionales con clara señal co-evolutiva en cada especie (según *CMG*) muestra que existe un elevado grado de especificidad en los pares de proteínas que han co-evolucionado en los linajes de diferentes especies. Sin embargo, un análisis de los procesos donde estas proteínas están implicadas muestra una sobre-representación de unos pocos procesos generales a lo largo de muchas de las especies analizadas. Estos procesos incluyen la cadena de transporte de electrones, la maquinaria flagelar y el transporte de sustratos a través de la membrana celular, así como procesos metabólicos de organofosfatos y amidas. Dichos resultados muestran que si bien algunos procesos parecen concentrar los fenómenos de co-evolución más fuertes, los pares concretos tienden a diferir entre especies. Considerando los procesos implicados, la situación descrita sugiere que estos fenómenos de co-evolución podrían estar implicados en los procesos de adaptación al medio de las diferentes especies.

Un análisis detallado de los procesos más representados en diferentes especies entre las proteínas que co-evolucionan, confirma que las interacciones funcionales que co-evolucionan difieren entre especies. Esta especificidad de especie, hace que la combinación de predicciones obtenidas para diferentes especies proporcione una imagen muy completa de dichos procesos. No obstante, este hecho no es impedimento para detectar que en sistemas complejos, como el flagelo o la NADH deshidrogenasa, la co-evolución es más común dentro de módulos funcionales o estructurales y entre módulos en contacto o espacialmente próximos. De hecho, es sorprendente el número relativamente bajo de casos de co-evolución detectados entre distintos complejos con funciones relacionadas. Esto sugiere que los complejos estructurales son la unidad de molecular donde más comúnmente se dan procesos de co-evolución, más habitualmente asociados a interacciones físicas.

4.3. Limitaciones y posibilidades futuras de *ContextMirror* y *ContextMirror Global*

4.3.1. Limitaciones y posibilidades futuras en el estudio de la co-evolución entre proteínas en procariotas

Las principales limitaciones de estas nuevas metodologías están asociadas a los posibles problemas para gestionar escenarios evolutivos complejos como las duplicaciones o los eventos de fusión génicas. Estos problemas son particularmente importantes en la aplicación de estos métodos a especies eucariotas, que se discute más abajo. Así mismo, cualquier fenómeno que altere la historia evolutiva de grupos pequeños de genes por motivos no funcionales puede generar parecidos en los árboles no relacionados con co-evolución, pero específicos de algunos genes. Aquí se discutirá el caso de los eventos de transferencia horizontal por su relevancia en procariotas, pero lo expuesto es aplicable a otros fenómenos de este tipo, como por ejemplo para el reparto incompleto de linaje.

A pesar de estas limitaciones los resultados presentados en esta tesis proporcionan una base sólida para la aplicación de *CM* y *CMG* a especies procariotas. Quizá con la única condición de tener suficientes especies evolutivamente próximas pero no redundantes. Es más, estos resultados dejan abiertas algunas preguntas que podrían ser objeto de futuras líneas de investigación y que también se discutirán.

4.3.1.A Eventos de transferencia horizontal

Las transferencias horizontales son comunes en procariotas por lo que pueden derivar en parecidos en los árboles de las proteínas que se han transferido juntas. De hecho en uno de los trabajos realizados durante esta tesis se observó que estos eventos contribuían negativamente a la detección de interacciones por *Tol-MirrorTree* (Pazos *et al.*, 2005), el antecedente de *CM* y *CMG*. *CM* y *CMG* no realizan una detección explícita de los eventos de transferencia horizontal, aunque el paso de *Perfiles Co-evolutivos* tiende a favorecer la detección de señales coherentes con el resto de proteínas del proteoma. Sin embargo, es posible que la aplicación en el futuro de un filtro específico de eventos de transferencia horizontal pudiera contribuir a mejorar los resultados obtenidos.

La introducción de la estandarización robusta en *MirrorTree* dentro del protocolo de *CMG*, podría facilitar la introducción de esta mejora. Dicha estandarización transforma las distancias evolutivas en desviaciones absolutas medianas de las distribuciones de distancias entre ortólogos del mismo par de especies. Eventos de transferencia horizontal entre linajes diferentes resultarán en desviaciones absolutas elevadas, por lo que un análisis de éstas debería identificar dichos casos. Este procedimiento sería análogo al planteado en *Tol-MirrorTree* (Pazos *et al.*, 2005).

Una vez detectados los casos de transferencia horizontal, se deberían explorar las ventajas y desventajas de eliminar estas proteínas del análisis *a priori* o etiquetarlas para su reconsideración *a posteriori*. Por un lado, excluir proteínas del análisis eliminaría la información necesaria para detectar su posible influencia en otros árboles, pero también podría reducir las distorsiones provenientes de introducir árboles muy diferentes del resto. Por tanto, no se debe descartar la posibilidad de que la introducción de ésta u otras correcciones en nuestros métodos pudiera mejorar su capacidad predictiva en el futuro.

4.3.1.B. ¿Cuál es el límite de la predicción basada en co-evolución?

Cuando se comenzó esta tesis los límites establecidos para los métodos basados en co-evolución estaban muy restringidos a su aplicación a casos concretos o la necesidad de lidiar con niveles elevados de predicciones erróneas (Pazos & Valencia, 2001). Los resultados presentados en esta memoria no sólo han mejorado notablemente la calidad de las predicciones, sino que han demostrado que la co-evolución es un fenómeno que afecta a miles de interacciones funcionales y que es prevalente en todas las especies bacterianas estudiadas. También han mostrado que la co-evolución es más común en determinados procesos biológicos aunque afecta a diferentes interacciones funcionales en diferentes especies.

La complementariedad en las interacciones detectadas en diferentes especies abre la puerta a obtener redes de interacciones funcionales de referencia para procariotas por acumulación de las detectadas en diferentes especies. Como se ha ilustrado con diferentes sistemas moleculares concretos esta aproximación tiene el potencial de proporcionar información más completa y detallada acerca de sus relaciones funcionales. Sin embargo, aún queda por determinar si este proceso acabaría por recuperar suficientes predicciones acerca de procesos no tan bien representados como la fosforilación oxidativa. Esta posible línea de investigación requeriría de un tratamiento explícito de las relaciones evolutivas entre las especies analizadas y de la composición de sus proteomas.

Otra posible línea de ampliación de estas redes de interacciones funcionales es mediante la combinación de análisis realizados con grupos de especies con diferentes grados de divergencia respecto a la especie de referencia. Nuestros resultados muestran que esta aproximación también produce predicciones diferentes aunque igualmente acertadas. Además en un trabajo al que se contribuyó durante esta tesis (Herman *et al.*, 2011) se muestra que estas interacciones parecen ser de naturaleza diferente, con las interacciones más lábiles mejor representadas en análisis centrados en la evolución reciente del linaje de *E. coli*. Así, la aplicación de nuestros métodos a la detección de señales de similitudes entre árboles a diferentes grados de divergencia evolutiva podría proporcionar información acerca de diferentes procesos biológicos, cuya importancia tenga diferente recorrido histórico. Si éste fuera el caso, se podría empezar por completar las redes de interacciones funcionales específicas de especies, para pasar después a integrar redes procedentes de especies diferentes.

En este contexto será también interesante ver como afecta el incremento en el número de especies disponibles. Como comentamos previamente los números de especies empleados en nuestros análisis están muy por debajo de lo adecuado para este tipo de metodologías. Por ello parece lógico esperar importantes mejoras asociadas a la incorporación de nuevas especies. No obstante, los excelentes resultados obtenidos en nuestros análisis sugieren que la estructura subyacente de la co-evolución entre proteínas puede ser más sencilla que la de otros sistemas relacionados. Esto implicaría que el éxito de estos métodos no estaría sólo asociado a la eliminación de correlaciones indirectas, sino a su capacidad de filtrar fenómenos inespecíficos de similitud entre grupos de proteínas, tales como la influencia del árbol de las especies o la co-evolución en grupos de distintos tamaños.

4.3.1.C. ¿Cuál es la importancia de la co-evolución en grupo?

Los resultados obtenidos para niveles no completamente específicos de *CM*, son compatibles con la existencia de una jerarquía de co-evoluciones asociada a la

co-evolución en grupo. Es más, nuestros resultados sugieren que esta jerarquía parece estar relacionada con diferentes niveles de co-dependencia estructural y funcional dentro de complejos grandes. A partir de estos resultados se plantea la hipótesis de que el establecimiento de la jerarquía co-evolutiva pueda aportar información más allá de la predicción de interacciones y revelar la estructura subyacente de las co-dependencias funcionales.

Sin embargo, *CM* no es capaz de recuperar dicha jerarquía explícitamente, ni de proponer un nivel óptimo de especificidad para cada par o grupo de proteínas. Es evidente que el descubrimiento de dichas jerarquías y establecer la relevancia de la co-evolución en grupo requeriría de nuevos desarrollos metodológicos que fueran capaces de discriminar la señal específica de grupo de la asociada a pares y de la inespecífica.

En este sentido son interesantes los esfuerzos realizados para modelar otros sistemas, como los patrones de respuesta de redes neurales. En estos sistemas se está explorando la incorporación de correlaciones de orden superior (en grupo) a modelos estadísticos del comportamiento de redes pequeñas (de hasta 100 nodos), aunque avances recientes sugieren que podrían alcanzarse tamaños mayores (Ganmor *et al.*, 2011; Tkacik *et al.*, 2013; Köster *et al.*, 2014). Algunas de estas aproximaciones están empezando a emplearse en la co-evolución a nivel de residuos con buenos resultados (Feinauer *et al.*, 2014; Contini & Tiana, 2015). En el futuro será interesante realizar un análisis profundo de los resultados de la aplicación de estos métodos a la co-evolución entre residuos, así como de la posible aplicación de éstas u otras estrategias relacionadas al problema de la co-evolución entre proteínas.

4.3.1.D. ¿Cuál es la relación entre co-evolución y adaptación evolutiva?

Los análisis realizados con *CMG* en 23 especies bacterianas, muestran que cada linaje evolutivo presenta señales de co-evolución entre diferentes proteínas. Sin embargo, existen procesos biológicos más habitualmente afectados por dichas señales. Estos procesos incluyen la obtención de energía, el intercambio de sustratos con el medio y la maquinaria flagelar bacteriana. Todos estos son procesos objeto de presiones de selección que varían en función del medio y por lo tanto están sometidos a constante adaptación. Esto abre la puerta a considerar la posibilidad de que las señales de co-evolución detectadas reflejen fenómenos de adaptación coordinada de las proteínas funcionalmente dependientes.

De hecho, los parecidos en los árboles de proteínas deben recoger, al menos parcialmente, señales de cambios simultáneos en la presión de selección de las proteínas que interaccionan funcionalmente. Esto refuerza la idea de que las señales detectadas puedan ser consecuencia de respuestas coordinadas de estas proteínas a cambios en las condiciones de vida de las especies en el linaje estudiado. Dicha respuesta coordinada podría incluir tanto fenómenos de co-adaptación, como fenómenos de presión de selección compartida como una misma unidad funcional.

En las definiciones más recientes de co-evolución entre especies se resalta la reciprocidad de los cambios entre los agentes evolutivos (ver Introducción). Esto derivó en una interpretación de la co-evolución a nivel de residuos en términos de cambios compensatorios orientados a conservar un contacto o interacción concreta (Korber *et al.*, 1993; Göbel *et al.*, 1994; Neher, 1994; Shindyalov *et al.*, 1994; Taylor & Hatrick, 1994). Sin embargo, el concepto de cambio adaptativo quedó casi excluido del

de co-evolución entre residuos, creando una separación artificial entre co-evolución y adaptación.

En el desarrollo de esta tesis, se ha defendido la opinión de que la reciprocidad es una idea bien representada por el concepto de co-adaptación, pero que el concepto de co-evolución debería recuperar su acepción más inclusiva, propuesta originalmente por Ehrlich y Raven (Ehrlich & Raven, 1964) en términos de "interacción evolutiva". Este cambio conceptual traería más claridad al campo de la co-evolución a nivel molecular, donde a menudo se confunden co-adaptación e interacción evolutiva. Así mismo, permitiría acoger con mayor naturalidad observaciones de co-dependencia evolutiva sin caer en sobre-interpretaciones acerca del proceso responsable de dicha co-dependencia (Juan *et al.*, 2008a; 2013).

4.3.2. Problemas y posibles estrategias futuras para la aplicación de *CM* y *CMG* en eucariotas

La aplicación de *CM* y *CMG*, como la de otros métodos basados en co-evolución, en eucariotas es aún problemática y supone un gran reto. Las características de la evolución de las proteínas en estas especies plantean desafíos importantes a las bases conceptuales y metodológicas de dichos métodos. Estos conceptos desafiados incluyen la definición de proteoma como sistema coherente en evolución, la asociación misma entre similitud de árboles y co-evolución o la definición de las proteínas como unidades básicas de evolución.

4.3.2.A. La variabilidad en la composición de los árboles de proteínas es mayor en eucariotas

A pesar de las ventajas discutidas previamente, analizar la co-evolución entre proteínas en el contexto del proteoma completo también presenta algunos problemas y limitaciones. Quizás el más importante de los cuales es el hecho evidente de que la composición del proteoma cambia a lo largo de la evolución. Estudiar la co-evolución en el proteoma de una especie, significa estudiar la red de parecidos entre los árboles de los ortólogos de sus proteínas en un conjunto de especies dado. Sin embargo, cada proteína de la especie de referencia presentará ortólogos en un subconjunto diferente de especies. Lógicamente, esto genera distorsiones en la red de correlaciones entre árboles, que tanto *CM* como *CMG* tratan de mitigar con diferentes estrategias. Estas estrategias funcionan correctamente en especies bacterianas cuando se incluyen un número razonable de especies en el análisis (más de 25 especies).

Sin embargo, el problema de la variabilidad en la composición de los árboles se acentúa en el estudio de genomas eucariotas. La mayor complejidad y frecuencia de los eventos de duplicación, pérdida, fusión y fisión de genes genera árboles de ortólogos únicos con distribuciones de especies extremadamente diferentes. En este contexto, la búsqueda de ortólogos no permite recuperar suficiente información, ni ésta es lo bastante consistente a lo largo de las proteínas de un proteoma. Así, las estrategias propuestas para corregir el efecto de las distorsiones en la correlaciones entre árboles se muestran insuficientes en el contexto del análisis de especies eucariotas.

Es esperable que el incremento en el número de genomas secuenciados palien parcialmente los efectos de este problema, pero, como se discute más adelante, también es necesario proporcionar una representación más equilibrada de la evolución eucariota. A pesar de todo, es posible que la aplicación de *CM* y *CMG* a eucariotas deba recurrir a definir preguntas más concretas que la co-evolución de todo el proteoma, o al menos a una resolución por partes de esta pregunta, como se discutirá posteriomente.

4.3.2.B. Conversión génica entre proteínas parálogas en especies con recombinación genética

La presencia de multitud de secuencias parálogas en los genomas eucariotas aumenta la importancia de los fenómenos de conversión génica entre homólogos. En la conversión génica entre homólogos, la recombinación de secuencias homólogas durante la mitosis resulta en el intercambio de material genético entre parálogos similares en secuencia. Estos fenómenos de conversión génica son difíciles de tratar en el contexto de nuestros métodos, porque implican que los árboles de parálogos no son realmente independientes entre sí, al menos durante un tiempo después de su duplicación. Una de las consecuencias de este efecto es la dificultad (y puede que imposibilidad) de discriminar señales co-evolutivas asociadas a parálogos recientes.

El tratamiento del problema de la conversión génica entre parálogos probablemente requeriría de estrategias específicas para evaluar este efecto y seleccionar proteínas y/u ortólogos cuya historia evolutiva sea razonablemente independiente (cuya especiación sea bastante posterior a las duplicaciones de los parálogos incluidos). Estas estrategias se pueden considerar análogas a las de selección de especies no redundantes, pero en este caso implicarían la selección de proteínas no redundantes.

4.3.2.C. Barajado de dominios entre proteínas

Otro problema que surge para la aplicación de *CM* y *CMG* en especies eucariotas es la definición de la unidad evolutiva adecuada. Es decir, mientras que en procariotas el estudio de la co-evolución entre proteínas parece ser una pregunta válida, en eucariotas, debido a los procesos de barajado de dominios, la pregunta más adecuada podría ser la co-evolución entre dominios de proteínas.

En este punto, es importante recalcar que la estrategia utilizada en esta memoria es semejante a la de estudiar aquellos dominios que representan la mayor parte de la proteína y permanecen presentes en los diferentes ortólogos, despreciando el efecto de los dominios que cambian. Esto implica, que una correcta interpretación de los análisis realizados en eucariotas deberá tener en cuenta qué regiones de cada proteína se están considerando. Por ejemplo, si los dominios implicados en la interacción de dos proteínas no están presentes en el resto de ortólogos, es improbable que se detecte la co-evolución asociada a la interacción.

Recuperar la información asociada a estos dominios móviles en estos datos requeriría de estrategias específicas para determinar qué dominios deberían estudiarse juntos (por ejemplo aquellos que permanecen en la mayoría de los ortólogos) y cuales por separado. Así mismo se podría detectar si la incorporación de un nuevo dominio implica una duplicación que requiera considerar los efectos de conversión génica discutidos anteriormente. Estos posibles desarrollos futuros podrían precisar de una integración de la información procedente de los árboles de proteínas y de los de dominios.

4.3.2.D. Problemas con los genomas eucariotas secuenciados

La combinación del problema del barajado de dominios y de la conversión génica, hacen del estudio de la co-evolución entre proteínas de eucariotas un gran reto, que deberá venir acompañado de la secuenciación de más genomas eucariotas que supongan una representación menos sesgada de la evolución eucariota. Los genomas secuenciados actualmente están muy sesgados hacia mamíferos, moscas y levaduras, lo

que también limita nuestras posibilidades de realizar una selección de genomas adecuada semejante a las presentadas en esta memoria para diferentes bacterias.

De hecho, actualmente es muy difícil recuperar un número de especies eucariotas con niveles de redundancia entre sí similares a los empleados en procariotas (salvo tal vez para levaduras). Además, esta reducción en el número de especies resulta mucho más problemática en eucariotas, ya que la alta frecuencia de duplicaciones y pérdidas de genes hace más difícil recuperar suficientes ortólogos únicos incluso cuando en teoría se disponga de suficientes especies.

Un problema relacionado es la relativamente baja calidad de muchos de los genomas eucariotas disponibles. Esto, unido a la mayor dificultad para predecir genes en especies eucariotas, nos sitúa en un escenario muy ruidoso. Por ejemplo, si la secuencia de los genes no está completamente resuelta o incluso no se puede predecir su presencia, la detección de ortólogos y la construcción de árboles se verá comprometida, reduciendo considerablemente la relación señal/ruido.

En resumen, las aproximaciones basadas en *MT*, se verán muy beneficiadas por el esperable incremento futuro de la cantidad y calidad de los genomas eucariotas secuenciados, aunque la medida de dicho beneficio dependerá también de la relación filogenética de estos nuevos genomas con los ya disponibles.

4.3.2.E. Análisis enfocados en detectar co-evolución en un número reducido de proteínas eucariotas

A pesar de la posible introducción de mejoras orientadas a tratar con especies eucariotas, la naturaleza de la evolución en estas especies impone grandes obstáculos a los análisis globales del proteoma. Una posible alternativa de futuro podría involucrar una táctica de "divide y vencerás". Es decir, se podrían diseñar estrategias para detectar co-evolución alrededor de una proteína (o un grupo reducido de proteínas). Esta proteína actuaría a modo de cebo, de forma similar a los diseños experimentales a gran escala de dos híbridos en levadura (Uetz *et al.*, 2000) o del aislamiento de complejos (Gavin *et al.*, 2002). Así, la selección de especies y la de qué proteínas se incluyen en el análisis estarían enfocadas a determinar los pares de co-evolución de esta proteína (o proteínas) cebo. Esta aproximación permitiría reducir las enormes dificultades para obtener una señal evolutiva coherente y robusta para un proteoma eucariota completo.

De esta forma, el foco se centraría en detectar señales asociadas a grupos de proteínas con distribuciones de especies semejantes, relacionadas funcionalmente y/o con correlaciones entre sus árboles más claras. Esto no implica necesariamente incluir pocas proteínas en el análisis, sino seleccionarlas de tal forma que permitan recuperar con mayor claridad las proteínas (presas) que co-evolucionan con la proteína cebo. Al igual que en el caso de dos híbridos en levadura, la aplicación sistemática de esta táctica nos podría permitir recuperar una visión más completa de red de co-evolución en eucariotas y con mayor precisión.

Como parte de un trabajo aún no publicado se ha dado un primer paso en esta dirección. En este trabajo se estudia la co-evolución entre 58 proteínas relacionadas con regulación epigenética y remodelación de la cromatina en *Mus musculus* (ratón). Los resultados obtenidos en este trabajo muestran que esta alternativa posibilita obtener

relaciones co-evolutivas de mayor calidad y nos ha permitido establecer una conexión interesante entre co-evolución y comunicación epigenética[‡] (Juan *et al.*, 2015).

4.4. Algunas reflexiones finales sobre la co-evolución en proteínas

En la última década se ha producido un auge de la aplicación de los principios de los sistemas complejos a la biología (Chuang *et al.*, 2010; Pujol *et al.*, 2010; Cox & Mann, 2011; Hogenesch & Ueda, 2011; Pritchard & Birch, 2011; Kandel *et al.*, 2014; Werner *et al.*, 2014). Esto no es de extrañar, ya que los organismos son comúnmente considerados como sistemas complejos organizados en una red de interacciones moleculares entre componentes básicos: proteínas, ADN, ARN, lípidos, azúcares, y otros (Joyce & Palsson, 2006). Este marco conceptual subraya la importancia de la red de interacciones entre los elementos que constituyen el sistema. Es decir, la importancia de los nodos (y sus propiedades), pero también de las interacciones entre ellos, junto a la dinámica de dichas interacciones. Así mismo, la comprensión de la naturaleza compleja de los sistemas biológicos conlleva la necesidad de abordar análisis globales en los que se pueda observar el comportamiento de todo el sistema. En este contexto, sólo desde una perspectiva holística será posible comprender los fenómenos evolutivos vinculados a la conservación y adaptación de los sistemas biológicos, y en particular de los relacionados con la co-evolución.

El trabajo presentado en esta memoria comparte este punto de vista al considerar al proteoma como un sistema complejo compuesto por una red de interdependencias funcionales entre proteínas. Así, se afronta la detección de señales co-evolutivas asociadas a estas interdependencias desde una estrategia "de arriba abajo". Es decir, se aborda el problema global para entender mejor cada uno de sus componentes. Los resultados obtenidos muestran el potencial de esta estrategia en la predicción de interacciones funcionales entre proteínas.

4.4.1. Las proteínas no son los únicos elementos que co-evolucionan, ni lo hacen de forma aislada

A pesar del avance que supone la visión de que el proteoma es "el sistema" y las proteínas "los elementos" que interaccionan, ésta es una decisión práctica que podría reconsiderarse en el futuro. De hecho, las proteínas no son los únicos elementos de los sistemas complejos vivos. Algunos de estos otros elementos también están incluidos en el genoma y están sometidos a procesos evolutivos (y de co-evolución), como los genes ARN no codificantes de proteínas, las regiones promotoras, las regiones reguladoras de la estructura del genoma, etc. Mientras que otros elementos son moléculas producidas por estos elementos o recuperadas del medio, que no están sometidas a evolución, pero que pueden actuar como intermediarias de las co-dependencias funcionales entre elementos que sí lo hacen.

Por ejemplo, en el trabajo mencionado anteriormente sobre comunicación epigenética, los resultados obtenidos también sugieren que la co-evolución puede darse entre

[‡] La comunicación implica el flujo de información entre un emisor y un receptor. Dicho flujo de información supone una interacción entre ambos agentes, que normalmente se manifiesta en un cambio de estado en el receptor. De hecho, muchos de los procesos co-evolutivos conllevan el establecimiento, adaptación y conservación de estrategias comunicativas. Es interesante que la adaptación de la señal y su efecto en el receptor se hayan invocado como criterios para definir una comunicación biológica real (Smith & Harper, 2003; Scott-Phillips, 2008).

proteínas conectadas por su tendencia a unirse o no a diferentes modificaciones de histonas o de citosinas (Juan *et al.*, 2015). Estos resultados ponen de manifiesto la necesidad de avanzar en una concepción aún más global del sistema en evolución, hasta una descripción molecular de la célula y en último término del organismo. En esta concepción, la interpretación de la co-evolución entre los elementos que sí evolucionan debería incluir el resto de elementos evolutivamente pasivos, pero cuya presencia o ausencia puede ser clave para entender la red de co-dependencias funcionales.

4.4.2. Los diferentes niveles evolutivos no son independientes entre sí

A lo largo de la introducción de esta tesis se ha hecho lo posible por presentar una visión multinivel del estudio de la co-evolución. Este esfuerzo obedece a una doble motivación. En primer lugar, se ha considerado necesario reflejar el papel esencial que el marco conceptual de la co-evolución entre especies juega en el desarrollo de metodologías para el estudio de niveles moleculares. En segundo lugar, y aún más importante, se ha pretendido recalcar que estos distintos niveles evolutivos no pueden comprenderse completamente de forma aislada. Esta visión, planteada desde la biología molecular (Juan *et al.*, 2013), ha sido también recientemente defendida desde el campo de la ecología evolutiva (Carmona *et al.*, 2015).

Hasta ahora, los métodos para detectar co-evolución han lidiado con esta situación tratando de eliminar la influencia de unos niveles sobre otros. Por ejemplo, los métodos presentados en esta memoria tratan de eliminar el efecto de la evolución de las especies o de grupos de secuencias, mientras que los métodos que trabajan a nivel de residuos hacen lo propio con las señales de las secuencias completas (Jones *et al.*, 2012; Ekeberg *et al.*, 2013). Sin embargo, estas aproximaciones no son capaces de aprovechar la información de los otros niveles para mejorar sus resultados y ayudar a su interpretación.

Así mismo, a lo largo de los trabajos realizados durante esta tesis, en los niveles de residuos y de proteínas, se han observado indicios de la relevancia de la co-evolución en grupo. Esta co-evolución en grupo podría suponer un estado intermedio entre la interacción entre elementos y el sistema que los engloba. La búsqueda de otras estructuras que engloben a grupos de elementos, como las de co-co-evolución (grupos de pares de elementos que co-evolucionan en las mismas condiciones), podrían ser líneas de investigación muy interesantes en el futuro que permitieran establecer conexiones entre los fenómenos evolutivos a diferentes niveles.

Quizás, una aproximación inspiradora en este contexto sean las metodologías dedicadas a la detección de *SDPs* (Casari *et al.*, 1995; Lichtarge *et al.*, 1996; del Sol Mesa *et al.*, 2003; Rausell *et al.*, 2010). Estas aproximaciones buscan señales a nivel de residuos que reflejan patrones informativos de las relaciones entre las secuencias de la familia. Estos patrones son indicativos de fenómenos de co-evolución en grupo con importantes consecuencias en el comportamiento del sistema (cambios en la función de las proteínas de la familia).

En este escenario parece que el desarrollo de un marco metodológico y conceptual capaz de abordar análisis co-evolutivos inter-nivel podría aportar una comprensión del fenómeno global de la co-evolución, así como de cada caso en particular. Este conocimiento, sin duda, permitiría un mejor aprovechamiento de la información evolutiva y conllevaría mejoras considerables en la capacidad predictiva de los métodos basados en co-evolución. Por lo tanto, puede que el desarrollo de metodologías capaces de analizar varios niveles simultáneamente, junto con la inclusión de métodos capaces

de detectar co-evolución en grupo de especies/proteínas/residuos estén entre los retos futuros más interesantes del estudio de la co-evolución.

Es evidente, que el campo del estudio de la co-evolución a nivel molecular es aún joven y que aún existen muchas incógnitas sobre los límites de su capacidad para reflejar los procesos moleculares que se observan en los organismos vivos. Sin embargo, ya ha producido multitud de resultados interesantes y de utilidad en la predicción de diferentes características moleculares. Además, estos resultados plantean numerosas preguntas nuevas, cuya exploración puede derivar en descubrimientos conceptuales con impredecibles implicaciones prácticas.

5. Conclusiones

1. Analizar las similitudes entre árboles de proteínas a nivel del proteoma completo, mediante la metodología adecuada de "mirror-trees", permite predecir interacciones funcionales entre proteínas con un alto nivel de confianza.
2. Es posible detectar con alta fiabilidad una señal asociada a la co-evolución de proteínas en grupos de proteínas que colaboran en funciones concretas.
3. La co-evolución en grupos de proteínas refleja aspectos de la organización estructural y funcional de complejos macromoleculares bien definidos.
4. También es posible detectar una señal adicional asociada a la co-evolución específica de pares de proteínas que está fuertemente asociada a su interacción física y funcional.
5. Las predicciones basadas en la detección co-evolución específica en diferentes especies bacterianas recuperan interacciones funcionales con alta fiabilidad y robustez.
6. La frecuencias de detección de co-evolución específica entre pares de proteínas ortólogas en diferentes especies bacterianas reflejan las relaciones estructurales y funcionales de los complejos moleculares, en especial los asociados con la obtención de energía, el intercambio de sustratos con el medio y la estructura flagelar.
7. Existe una relación entre el número, tipo y distribución de secuencias de proteínas de los distintos organismos disponibles para el cálculo de señal co-evolutiva y la fiabilidad de las interacciones predichas.
8. La señal de co-evolución específica y la detectada en la co-evolución de grupos de proteínas, permiten detectar interacciones funcionales diferentes, aunque la co-evolución específica muestra mayor capacidad predictiva.
9. A nivel general, los métodos basados en la similitud de árboles filogenéticos, representan un área de desarrollo particularmente interesante en el campo de la co-evolución molecular. En el futuro, será necesaria la integración de estos métodos con aquellos basados en el análisis de los patrones de sustitución en posiciones de alineamientos múltiples.

6. Materiales y Métodos

Como se explica en la sección de Resultados esta tesis presenta el desarrollo y los análisis realizados con dos nuevas metodologías para la detección de co-evolución entre pares de proteínas. Por lo tanto, al igual que en la sección de Resultados, se organizará la sección de Materiales y Métodos en dos sub-secciones referentes a cada una de las metodologías desarrolladas.

6.1. Materiales y métodos asociados a los análisis realizados con *ContextMirror*

6.1.1. Base de datos de genomas secuenciados y selección de genomas

6.1.1.A. Integr8 (Kersey et al., 2005) (http://www.ebi.ac.uk/integr8/)

Integr8 era una base de datos que integraba información acerca de genomas secuenciados y de sus correspondientes proteomas. Esta información provenía originalmente de otras bases de datos tales como *EMBL Nucleotide Sequence Database* (Kulikova *et al.*, 2007), *ENSEMBL* (Flicek *et al.*, 2008), *UniProt* (UniProt Consortium, 2007) o *IPI* (Kersey *et al.*, 2004). Cada una de las cuales son recursos de referencia para el almacenamiento y organización de diferentes aspectos de las secuencias genómicas, génicas y proteicas. Integr8 era un recurso particularmente útil para análisis a gran escala, ya que permitía relacionar de forma sencilla información genómica y proteómica. La información referente a las secuencias de proteínas de los genomas utilizados en los análisis con *ContextMirror* fueron obtenidos de Integr8.

Durante el desarrollo de esta tesis, Integr8 fue sustituido por *ENSEMBL Genomes* (Kersey *et al.*, 2010), que supone una extensión del previamente existente *ENSEMBL* (Flicek *et al.*, 2011) a genomas de organismos no vertebrados.

6.1.1.B. Selección de genomas

La selección de genomas se realizó a partir de los 218 genomas completamente secuenciados de procariotas disponibles en Integr8 en febrero del 2005. Con intención de reducir la redundancia de estos genomas se utilizó el árbol taxonómico del NCBI (Federhen, 2012). Se seleccionó una especie de cada grupo de especies relacionadas en el último nivel del árbol taxonómico, que, en general, coincide con la clasificación de cepa. Para minimizar la pérdida de información, la especie con mayor número de proteínas fue seleccionada como representante de cada grupo. La aplicación de este proceso resultó en un conjunto de 114 especies procariotas (Tabla 2).

Tabla 2. Especies empleadas en el análisis de *CM* con *E. coli* como especie de referencia.

TaxID	Nombre de especie	Filo	Dominio
206672	*Bifidobacterium longum (strain NCC 2705)*	Actinobacteria	Bacteria
196627	*Corynebacterium glutamicum (strain ATCC 13032)*	Actinobacteria	Bacteria
281090	*Leifsonia xyli subsp. xyli (strain CTCB07)*	Actinobacteria	Bacteria
272631	*Mycobacterium leprae (strain TN)*	Actinobacteria	Bacteria

TaxID	Nombre de especie	Filo	Dominio
262316	*Mycobacterium paratuberculosis (strain ATCC BAA-968)*	Actinobacteria	Bacteria
83331	*Mycobacterium tuberculosis (strain CDC 1551)*	Actinobacteria	Bacteria
247156	*Nocardia farcinica (strain IFM 10152)*	Actinobacteria	Bacteria
267747	*Propionibacterium acnes (strain KPA171202)*	Actinobacteria	Bacteria
100226	*Streptomyces coelicolor (strain ATCC BAA-471)*	Actinobacteria	Bacteria
203267	*Tropheryma whipplei (strain Twist)*	Actinobacteria	Bacteria
180835	*Agrobacterium tumefaciens str. C58*	Alphaproteobacteria	Bacteria
234826	*Anaplasma marginale (strain St. Maries)*	Alphaproteobacteria	Bacteria
38323	*Bartonella henselae*	Alphaproteobacteria	Bacteria
224911	*Bradyrhizobium diazoefficiens (strain JCM 10833)*	Alphaproteobacteria	Bacteria
204722	*Brucella suis biovar 1 (strain 1330)*	Alphaproteobacteria	Bacteria
190650	*Caulobacter crescentus (strain ATCC 19089)*	Alphaproteobacteria	Bacteria
266835	*Rhizobium loti (strain MAFF303099)*	Alphaproteobacteria	Bacteria
266834	*Rhizobium meliloti (strain 1021)*	Alphaproteobacteria	Bacteria
258594	*Rhodopseudomonas palustris (strain ATCC BAA-989)*	Alphaproteobacteria	Bacteria
272944	*Rickettsia conorii (strain ATCC VR-613)*	Alphaproteobacteria	Bacteria
257363	*Rickettsia typhi (strain ATCC VR-144)*	Alphaproteobacteria	Bacteria
246200	*Ruegeria pomeroyi (strain ATCC 700808)*	Alphaproteobacteria	Bacteria
66077	*Wolbachia endosymbiont of Drosophila melanogaster*	Alphaproteobacteria	Bacteria
264203	*Zymomonas mobilis subsp. mobilis (strain ATCC 31821)*	Alphaproteobacteria	Bacteria
224324	*Aquifex aeolicus (strain VF5)*	Aquificae	Bacteria
226186	*Bacteroides thetaiotaomicron (strain ATCC 29148)*	Bacteroidetes / Chlorobi	Bacteria
194439	*Chlorobium tepidum (strain ATCC 49652)*	Bacteroidetes / Chlorobi	Bacteria
242619	*Porphyromonas gingivalis (strain ATCC BAA-308)*	Bacteroidetes / Chlorobi	Bacteria
76114	*Aromatoleum aromaticum (strain EbN1)*	Betaproteobacteria	Bacteria
518	*Bordetella bronchiseptica*	Betaproteobacteria	Bacteria
243160	*Burkholderia mallei (strain ATCC 23344)*	Betaproteobacteria	Bacteria
272560	*Burkholderia pseudomallei (strain K96243)*	Betaproteobacteria	Bacteria
243365	*Chromobacterium violaceum (strain ATCC 12472)*	Betaproteobacteria	Bacteria
122587	*Neisseria meningitidis serogroup A / serotype 4A (strain Z2491)*	Betaproteobacteria	Bacteria
228410	*Nitrosomonas europaea (strain ATCC 19718)*	Betaproteobacteria	Bacteria
267608	*Ralstonia solanacearum (strain GMI1000)*	Betaproteobacteria	Bacteria

TaxID	Nombre de especie	Filo	Dominio
272561	*Chlamydia trachomatis (strain D)*	Chlamydiae / Verrucomicrobia	Bacteria
182082	*Chlamydophila pneumoniae TW-183*	Chlamydiae / Verrucomicrobia	Bacteria
264201	*Protochlamydia amoebophila (strain UWE25)*	Chlamydiae / Verrucomicrobia	Bacteria
251221	*Gloeobacter violaceus (strain PCC 7421)*	Cyanobacteria	Bacteria
103690	*Nostoc sp. (strain PCC 7120)*	Cyanobacteria	Bacteria
74547	*Prochlorococcus marinus (strain MIT 9313)*	Cyanobacteria	Bacteria
84588	*Synechococcus sp. (strain WH8102)*	Cyanobacteria	Bacteria
1148	*Synechocystis sp. PCC 6803*	Cyanobacteria	Bacteria
1299	*Deinococcus radiodurans*	Deinococcus-Thermus	Bacteria
262724	*Thermus thermophilus (strain HB27)*	Deinococcus-Thermus	Bacteria
959	*Bdellovibrio bacteriovorus*	Deltaproteobacteria	Bacteria
177439	*Desulfotalea psychrophila (strain LSv54)*	Deltaproteobacteria	Bacteria
882	*Desulfovibrio vulgaris (strain Hildenborough)*	Deltaproteobacteria	Bacteria
243231	*Geobacter sulfurreducens (strain ATCC 51573)*	Deltaproteobacteria	Bacteria
192222	*Campylobacter jejuni subsp. jejuni serotype O:2 (strain NCTC 11168)*	Epsilonproteobacteria	Bacteria
235279	*Helicobacter hepaticus (strain ATCC 51449)*	Epsilonproteobacteria	Bacteria
273121	*Wolinella succinogenes (strain ATCC 29543)*	Epsilonproteobacteria	Bacteria
222523	*Bacillus cereus (strain ATCC 10987)*	Firmicutes	Bacteria
279010	*Bacillus licheniformis*	Firmicutes	Bacteria
273068	*Caldanaerobacter subterraneus subsp. tengcongensis (strain DSM 15242)*	Firmicutes	Bacteria
272562	*Clostridium acetobutylicum (strain ATCC 824)*	Firmicutes	Bacteria
226185	*Enterococcus faecalis (strain ATCC 700802)*	Firmicutes	Bacteria
235909	*Geobacillus kaustophilus (strain HTA426)*	Firmicutes	Bacteria
220668	*Lactobacillus plantarum (strain ATCC BAA-793)*	Firmicutes	Bacteria
272623	*Lactococcus lactis subsp. lactis (strain IL1403)*	Firmicutes	Bacteria
272626	*Listeria innocua serovar 6a (strain CLIP 11262)*	Firmicutes	Bacteria
265311	*Mesoplasma florum (strain ATCC 33453)*	Firmicutes	Bacteria
221109	*Oceanobacillus iheyensis (strain DSM 14371)*	Firmicutes	Bacteria
262768	*Onion yellows phytoplasma (strain OY-M)*	Firmicutes	Bacteria
158878	*Staphylococcus aureus (strain Mu50)*	Firmicutes	Bacteria
208435	*Streptococcus agalactiae serotype V (strain ATCC BAA-611)*	Firmicutes	Bacteria
2734	*Symbiobacterium thermophilum*	Firmicutes	Bacteria
190304	*Fusobacterium nucleatum subsp. nucleatum (strain ATCC 25586)*	Fusobacteria	Bacteria
62977	*Acinetobacter baylyi (strain ATCC 33305)*	Gammaproteobacteria	Bacteria

TaxID	Nombre de especie	Filo	Dominio
203907	*Blochmannia floridanus*	Gammaproteobacteria	Bacteria
107806	*Buchnera aphidicola subsp. Acyrthosiphon pisum (strain APS)*	Gammaproteobacteria	Bacteria
227377	*Coxiella burnetii (strain RSA 493)*	Gammaproteobacteria	Bacteria
83333	*Escherichia coli (strain K12)*	Gammaproteobacteria	Bacteria
71421	*Haemophilus influenzae (strain ATCC 51907)*	Gammaproteobacteria	Bacteria
283942	*Idiomarina loihiensis (strain ATCC BAA-735)*	Gammaproteobacteria	Bacteria
272624	*Legionella pneumophila subsp. pneumophila (strain Philadelphia 1)*	Gammaproteobacteria	Bacteria
221988	*Mannheimia succiniciproducens (strain MBEL55E)*	Gammaproteobacteria	Bacteria
243233	*Methylococcus capsulatus (strain ATCC 33009)*	Gammaproteobacteria	Bacteria
272843	*Pasteurella multocida (strain Pm70)*	Gammaproteobacteria	Bacteria
218491	*Pectobacterium atrosepticum (strain SCRI 1043)*	Gammaproteobacteria	Bacteria
74109	*Photobacterium profundum*	Gammaproteobacteria	Bacteria
243265	*Photorhabdus luminescens subsp. laumondii (strain TT01)*	Gammaproteobacteria	Bacteria
208964	*Pseudomonas aeruginosa (strain ATCC 15692)*	Gammaproteobacteria	Bacteria
601	*Salmonella enterica subsp. enterica serovar Typhi*	Gammaproteobacteria	Bacteria
211586	*Shewanella oneidensis (strain MR-1)*	Gammaproteobacteria	Bacteria
198214	*Shigella flexneri 2a str. 301*	Gammaproteobacteria	Bacteria
196600	*Vibrio vulnificus (strain YJ016)*	Gammaproteobacteria	Bacteria
36870	*Wigglesworthia glossinidia brevipalpis*	Gammaproteobacteria	Bacteria
190486	*Xanthomonas axonopodis pv. citri (strain 306)*	Gammaproteobacteria	Bacteria
160492	*Xylella fastidiosa (strain 9a5c)*	Gammaproteobacteria	Bacteria
273123	*Yersinia pseudotuberculosis serotype I (strain IP32953)*	Gammaproteobacteria	Bacteria
117	*Pirellula sp.*	Planctomycetes	Bacteria
224326	*Borrelia burgdorferi (strain ATCC 35210)*	Spirochaetes	Bacteria
189518	*Leptospira interrogans serogroup Icterohaemorrhagiae serovar Lai (strain 56601)*	Spirochaetes	Bacteria
243275	*Treponema denticola (strain ATCC 35405)*	Spirochaetes	Bacteria
272633	*Mycoplasma penetrans (strain HF-2)*	Tenericutes	Bacteria
273119	*Ureaplasma parvum serovar 3 (strain ATCC 700970)*	Tenericutes	Bacteria
243274	*Thermotoga maritima (strain ATCC 43589)*	Thermotogae	Bacteria
272557	*Aeropyrum pernix (strain ATCC 700893)*	Crenarchaeota	Archaea
13773	*Pyrobaculum aerophilum*	Crenarchaeota	Archaea
2287	*Sulfolobus solfataricus*	Crenarchaeota	Archaea
224325	*Archaeoglobus fulgidus (strain ATCC 49558)*	Euryarchaeota	Archaea

TaxID	Nombre de especie	Filo	Dominio
272569	*Haloarcula marismortui (strain ATCC 43049)*	Euryarchaeota	Archaea
64091	*Halobacterium salinarum (strain ATCC 700922)*	Euryarchaeota	Archaea
243232	*Methanocaldococcus jannaschii (strain ATCC 43067)*	Euryarchaeota	Archaea
39152	*Methanococcus maripaludis*	Euryarchaeota	Archaea
190192	*Methanopyrus kandleri (strain AV19)*	Euryarchaeota	Archaea
188937	*Methanosarcina acetivorans (strain ATCC 35395)*	Euryarchaeota	Archaea
187420	*Methanothermobacter thermautotrophicus (strain ATCC 29096)*	Euryarchaeota	Archaea
263820	*Picrophilus torridus (strain ATCC 700027)*	Euryarchaeota	Archaea
70601	*Pyrococcus horikoshii (strain ATCC 700860)*	Euryarchaeota	Archaea
273116	*Thermoplasma volcanium (strain ATCC 51530)*	Euryarchaeota	Archaea
228908	*Nanoarchaeum equitans (strain Kin4-M)*	Nanoarchaeota	Archaea

6.1.2. Obtención de la matriz de distancias filogenéticas para las proteínas de *E. coli*

6.1.2.A. Detección de ortólogos mediante BBH (del inglés Best Bi-directional Hit; Tatusov 1996)

Dos proteínas homólogas son ortólogas entre sí, si la escisión de sus historias evolutivas se debió a un evento de especiación (Fitch, 1970). Dicha escisión es improbable que diera lugar a una diferenciación funcional entre ambas proteínas, ya que es esperable que siguieran desarrollando en sus respectivas especies una función similar a la de su ancestro común. El escenario alternativo más común es cuando la escisión se debe a un evento de duplicación, dando lugar a relaciones de paralogía. Como se ha comentado en la introducción, una parte importante de la innovación funcional parece provenir de estos eventos de duplicación, que generan condiciones en que una o ambas proteínas resultantes pueden desarrollar funciones diferentes (Ohno, 1970; Force *et al.*, 1999; Stoltzfus, 1999; Lynch & Conery, 2000; Kondrashov *et al.*, 2002; Innan & Kondrashov, 2010).

Esta visión simplificada de la relación entre ortología/paralogía y conservación/innovación funcional no refleja la complejidad real del escenario evolutivo. En primer lugar, en ambos casos, en el momento de la escisión, las dos líneas evolutivas tienen una alta probabilidad de mantener la función original. Es con el paso del tiempo que se favorece la divergencia funcional de forma mucho más marcada en las paralogías. Por lo tanto, en igualdad de condiciones, es esperable que los parálogos diverjan más funcionalmente. Sin embargo, la situación se complica con la acumulación de eventos de duplicación posteriores a una especiación. En estos casos, los grupos de parálogos de uno de los linajes surgidos de la duplicación son ortólogos de los grupos de parálogos del otro. Dado que las duplicaciones posteriores han podido derivar en diferentes divergencias funcionales, estas relaciones de ortología tienen menos probabilidad de reflejar una equivalencia funcional.

En esta tesis se utiliza una definición de trabajo muy exigente de ortología, conocida como "el mejor resultado bidireccional" o *BBH* (del inglés *Best Bi-directional Hit*) (Tatusov *et al.*, 1996). Esta definición establece que dos secuencias son ortólogas entre sí, si y sólo si ambas son el mejor resultado encontrado en su proteoma para una búsqueda por homología con *BLAST* (Altschul *et al.*, 1997) iniciada con la otra secuencia. Esta estrategia, presenta varias ventajas sobre el uso de la definición clásica de ortología. En primer lugar, su sencillez comparada con la resolución de historias evolutivas a gran escala, un problema sin resolver y que requiere grandes recursos computacionales. En segundo lugar, esta definición está más próxima al ideal de conservación funcional, ya que pretende detectar proteínas funcionalmente equivalentes. En un contexto filogenético, esta definición permite detectar los pares de ortólogos con mayor probabilidad de ser funcionalmente equivalentes. Por último, permite simplificar el estudio de correlaciones en las historias evolutivas, ya que el número de secuencias por especie nunca es superior a uno (requisito para metodologías tipo *MT*).

Los principales problemas de esta definición están asociados a dos factores: las pérdidas génicas y el carácter local de los resultados de *BLAST*. Las pérdidas génicas son un problema cuando existe otro parálogo del gen perdido en su especie. Al desaparecer el gen ortólogo, nuestro criterio puede confundir al gen parálogo con el ortólogo de nuestra proteína de referencia. En la mayoría de los casos, no debería darse esta confusión, a no ser que en el genoma de referencia no exista un ortólogo del parálogo del gen perdido. Esta situación se verá atenuada por la tendencia de los genes a recuperar la función de sus genes parálogos perdidos (DeLuna *et al.*, 2008). Así mismo, esta situación es mucho menos común en genomas procariotas, donde la mayoría de los genes tienen relativamente pocos parálogos. No obstante, es esperable que este problema siga dándose en algunos casos.

El problema del carácter local de los resultados de *BLAST* implica que estas evaluaciones de ortología se hacen en virtud de las regiones detectadas como homólogas por *BLAST* (Altschul *et al.*, 1997). Estas regiones pueden ser bastante cortas, por lo que eventos de corte de genes o reorganización de dominios, podrían dar lugar a confusiones. Con la intención de reducir estos posibles problemas, Se estableció el criterio adicional de que ambas búsquedas de *BLAST* debían asociar un *e-value* menor de 10^{-5} a la detección del correspondiente posible ortólogo (habiendo fijado la longitud efectiva de la base de datos a 10^{-8}), para evitar la inclusión de homologías poco fiables. Además, se estableció el filtro de que ambas proteínas debían alinear más de un 70% de su secuencia con la otra, con lo que se evitaron problemas derivados de la detección de similitudes parciales de origen ambiguo, pero que en cualquier caso, no bastan para sugerir una equivalencia funcional. El protocolo de *BBH* fue implementado en la forma de scripts de Perl. Este protocolo se utilizó para obtener los ortólogos únicos de todas las proteínas de *Escherichia coli K12* en todas las especies procariotas seleccionadas (ver Tabla 2).

6.1.2.B Construcción de alineamientos múltiples.

Una vez recuperadas las relaciones de ortología para cada proteína de *E. coli* en los otros genomas, se seleccionaron aquellos grupos de ortólogos con más de 15 miembros (2.183 alineamientos). Posteriormente realizaron alineamientos múltiples de secuencias para cada uno de estos conjuntos de secuencias ortólogas. Estos alineamientos fueron realizados usando el software *MUSCLE* (Edgar, 2004). Este software tiene una

excelente relación entre calidad de alineamiento y eficiencia computacional, lo que es fundamental para un análisis a gran escala como el nuestro.

6.1.2.C Cálculos de árboles filogenéticos.

Los alineamientos múltiples de grupos de ortólogos se utilizaron para construir sus respectivos árboles filogenéticos. Para ello, se empleó el algoritmo de unión de vecinos (Saitou & Nei, 1987) implementado en *ClustalW* (Larkin *et al*., 2007). En concreto, se utilizó la corrección de distancias por substitución múltiple de Kimura (Kimura, 1980; Pazos *et al.*, 2005) y se eliminaron aquellas columnas con algún hueco. Los árboles obtenidos deben considerarse como una aproximación razonable de la historia evolutiva de estas proteínas en el contexto de un análisis a gran escala. Un análisis filogenético detallado de una familia de proteínas requiere de la toma de mucha decisiones manualmente, así como de un tiempo considerable, y por lo tanto estaba fuera de las posibilidades en estos análisis.

6.1.2.D Construcción de la matriz de distancias

Finalmente, se extrajeron de los árboles de ortólogos las distancias entre las proteínas de cada par de especies incluidas. Estas distancias se calcularon como la suma de las longitudes de las hojas del árbol correspondiente hasta el último ancestro común a ambas. Estas distancias se utilizaron para construir una matriz de distancias para cada árbol, también conocida como matriz cofenética (Sokal & Rohlf, 1962). Estas matrices de distancias fueron utilizadas para construir una matriz global con una fila para cada árbol (es decir para cada proteína de *E. coli*) y en la que cada columna corresponde a la distancia entre los ortólogos en dos especies concretas a lo largo de todos los árboles. En esta matriz, aquellas posiciones sin valor calculado (por falta de ortólogos para esa proteína en esas especies) se dejaron en blanco. La matriz resultante contenía 2.183 filas y 6.441 columnas y constituye la matriz de partida para los análisis realizados con *ContextMirror* (ver Sección 3.1. Co-evolución específica de grupos de proteínas.).

6.1.3 Construcción de los conjuntos de evaluación

6.1.3.A. Interacciones físicas entre proteínas (LT_PPI)

Para el análisis realizado con *ContextMirror*, se recopilaron las interacciones físicas entre proteínas de *E. coli*, basadas en experimentos a pequeña escala procedentes de las bases de datos: *DIP* (http://dip.doe-mbi.ucla.edu/) (Salwinski *et al.*, 2004); *BIND* (http://binddb.org) (Bader *et al.*, 2001); *MINT* (http://mint.bio.uniroma2.it/mint/) (Zanzoni *et al.*, 2002) e *INTACT* (http://www.ebi.ac.uk/intact/) (Kerrien *et al.*, 2007)**.** Todas éstas son bases de datos de amplio espectro en las que se almacenan interaccionen recuperadas a través de un proceso de revisión manual de la literatura científica. Estas bases de datos difieren en la información acumulada sobre la interacción, así como en la forma de organizar dicha información. Para todas las bases de datos se recuperaron los identificadores de proteína y la publicación en la que se detectó la interacción. Con el fin de seleccionar experimentos realizados a pequeña escala, se eliminaron todas aquellas interacciones cuya única evidencia provenía de un artículo que reportara más de 50 interacciones en total. La integración de estas bases de datos produjo el conjunto *LT_PPI*, que contiene 3.965 interacciones entre 812 proteínas.

6.1.3.B. Complejos entre proteínas

i) Complejos procedentes de *EcoCyc* (Keseler, 2004)

La base de datos *EcoCyc* contiene una colección de complejos entre proteínas de alta fiabilidad extraídos de la literatura manualmente y detectados mediante técnicas experimentales a pequeña escala. La vocación de este conjunto no es tanto recuperar un conjunto completo de complejos, como disponer de una definición de alta calidad de complejos caracterizados bioquímicamente con función conocida.

ii) Complejos procedentes de experimentos a gran escala

En el momento del desarrollo de los análisis con *ContextMirror*, se habían publicado dos experimentos de aislamiento de complejos a gran escala (Butland *et al.*, 2005; Arifuzzaman, 2006). Estos estudios utilizaron técnicas de purificación por afinidad en tándem seguidas de espectrometría de masas. De esta manera, es posible detectar conjuntos de proteínas que pertenecen al mismo complejo. Estas técnicas se han aplicado con buenos resultados en distintos sistemas, pero como todas las técnicas a gran escala, también se tienden a detectar falsos complejos (Mering *et al.*, 2002). En este caso se decidió mezclar los conjuntos de ambos experimentos con la intención de recuperar un imagen lo más completa posible de los complejos existentes en *E. coli*.

iii) Construcción de los conjuntos de evaluación *LT_COMPLEX* y *HT_COMPLEX*

Los conjuntos de complejos discutidos proporcionan información acerca de la pertenencia de las proteínas al mismo complejo, pero no acerca de su disposición espacial. La forma de traducir estos datos a pares de proteínas que se ha elegido es la que se denomina como la aproximación matriz (Bader & Hogue, 2002). Es decir, que se establecen pares de proteínas entre todos los miembros del complejo. Esta aproximación aunque incluye interacciones no físicas, recupera pares de proteínas con una co-dependencia funcional clara, ya que los complejos forman unidades funcionales en la que todos sus miembros contribuyen a la función del complejo. Mediante la aplicación de esta estrategia al conjunto de 245 complejos de *EcoCyc*, se obtuvo el conjunto *LT_COMPLEX*. *LT_COMPLEX* contiene 1.354 pares entre 591 proteínas. De forma similar, se construyó *HT_COMPLEX* a partir la combinación de experimentos de aislamiento de complejos a gran escala. *HT_COMPLEX* contiene 53.002 pares entre 2.842 proteínas. La diferencia entre ambos conjuntos evidencia las enormes diferencias en la generación de estos conjuntos y sus distintos objetivos.

6.1.3.C Rutas metabólicas

i) *EcoCyc* (Keseler, 2004)

EcoCyc es una base de datos que contiene información sobre rutas metabólicas y regulación génica para *Escherichia coli K12*. *EcoCyc* extrae la información manualmente de la literatura y la integra dentro de un entorno más amplio conocido como *BioCyc*. *BioCyc* contiene otras cinco bases de datos similares a *EcoCyc* para otras especies (*e.g. HumanCyc* para humanos) así como otra base de datos denominada *MetaCyc* que contiene predicciones revisadas manualmente para más 5.000 organismos diferentes, obtenidas aprovechando la información de las seis bases de datos anteriores y de otros recursos externos.

ii) *KEGG* (Kanehisa *et al.*, 2004)

KEGG es un recurso que contiene gran cantidad de información acerca de genes y genomas completos. En concreto, *KEGG Pathway* contiene información acerca de rutas metabólicas extendida a más de 4.000 organismos a través de sus predicciones de ortología. *KEGG* asigna funciones enzimáticas a los genes y de esta forma establece su situación en las rutas metabólicas definidas manualmente.

iii) Construcción de los conjuntos de evaluación *ECOCYC_PWY* y *KEGG_PWY*

El concepto de ruta metabólica está bien establecido en el campo como el conjunto de reacciones bioquímicas que conducen de uno o más substratos a uno o más productos, con la generación una serie de compuestos intermedios. Sin embargo, la aplicación de esta definición es inevitablemente ambigua, y está sujeta tanto a los convencionalismos del campo como a decisiones subjetivas. Por lo tanto, en este caso se decidió establecer dos conjuntos de asociaciones funcionales basados en rutas metabólicas obtenidos a partir de las bases de datos *EcoCyc* y *KEGG*.

EcoCyc y *KEGG* tienen una estructura interna muy diferente, a pesar de lo cual en ambos casos es posible recuperar la relación entre proteínas y rutas metabólicas en las que actúan. De esta manera, se pudieron definir en ambos casos relaciones entre cada par de proteínas anotadas en la misma ruta metabólica. Una virtud de esta aproximación es que no asume nada acerca de las relaciones entre los pares, por lo que permite observar cualquier co-evolución detectada en el contexto de una ruta dada.

Esta estrategia es particularmente relajada para el caso de las rutas metabólicas, ya que una ruta metabólica puede estar compuesta por cientos de proteínas con una dependencia funcional entre ellas muy variable. De hecho, se obtuvieron 78.532 pares entre 1.339 proteínas en *KEGG* (conjunto *KEGG_PWY*) y 4.491 pares entre 719 proteínas en *EcoCyc* (conjunto *ECOCYC_PWY*). Estos números muestran que si bien *KEGG* contiene más información, ésta es mucho menos específica. Dos factores contribuyen a estas diferencias. E n primer lugar *KEGG* contiene rutas de mayor tamaño. Además, *KEGG* usa funciones catalíticas inferidas en la asignación de proteínas a rutas. Esto significa que dos proteínas anotadas con la misma función, serán asignadas a todas las rutas que contengan dicha función, lo que claramente ignora la importancia de proteínas específicas de ruta, condiciones de concentración, etc.

6.1.3.D. Construcción de los conjuntos de negativos

Aunque la determinación de si dos proteínas tienen una determinada asociación funcional está sometida a ciertas incertidumbres asociadas a la fiabilidad de los datos de

partida, la determinación de cuando dos proteínas no interaccionan funcionalmente es aún más problemática. La ausencia de evidencia de una interacción no necesariamente implica que no exista. De hecho, ni siquiera implica que se haya intentado determinar si existe. Es más, las diferentes técnicas experimentales tienen sus propios sesgos en la detección de interacciones. Para reducir la influencia de estos sesgos, se establecieron conjuntos de negativos para cada conjunto de evaluación a partir de las asociaciones incluidas. Estos conjuntos de negativos se definieron como los pares de proteínas sin asociación en el conjunto de evaluación, pero en los que ambas proteínas tienen al menos una asociación en dicho conjunto. La virtud de esta aproximación es que la evaluación se realiza únicamente para aquellas proteínas para las que se sabe que se han hecho experimentos capaces de determinar alguna de sus asociaciones funcionales. La limitación de esta aproximación es que sigue asumiendo que se conocen todas las asociaciones funcionales de estas proteínas, lo que en muchos casos no es cierto.

Siguiendo esta estrategia, el número de negativos de cada conjunto se puede determinar a partir de los pares de proteínas en el conjunto como:

$$Neg = \left(\frac{n \times (n-1)}{2} \right) - Pos ,$$

donde *n* corresponde al número de proteínas en el conjunto y *Pos* al número de asociaciones funcionales el mismo.

Por ejemplo, *LT_COMPLEX* que tiene 1.354 asociaciones entre 591 proteínas tendrá 172.991 pares negativos entre esas proteínas. Así, se obtuvieron 325.301 negativos para *LT_PPI*, 3.986.901 para *HT_COMPLEX*, 817.259 para *KEGG_PWY* y 253.630 para *ECOCYC_PWY*.

6.1.4. Evaluación de las predicciones obtenidas por *ContextMirror*

6.1.4.A. Evaluación de la precisión predictiva de ContextMirror

Primero se realizó un análisis detallado de la precisión obtenida para las mejores predicciones de la metodología desarrollada en los distintos conjuntos de evaluación. Para lo cual, se calculó la precisión incremental a lo largo de las *N* primeras predicciones (con $100 \leq N \leq 2.000$. La precisión es definida como:

$$Precisión = \frac{TP}{TP + FP} ,$$

donde *TP* corresponde al número de casos detectados que tienen una asociación funcional y *FP* al número de casos que no la tienen en el conjunto dado. Esta precisión se calcula para las *N* primeras predicciones, de las que se evalúan aquellas presentes en el correspondiente conjunto de evaluación o en su respectivo conjunto de negativos.

6.1.4.B. Evaluación de la capacidad discriminativa global (análisis ROC)

La evaluación de capacidad discriminativa de un método requiere de una evaluación global tanto de su especificidad, como su sensibilidad. Una de las mejores formas de realizar esta evaluación es mediante un análisis *ROC* (del inglés *Receiver Operating Characteristic*). Este análisis consiste en representar de forma recursiva la *Sensibilidad* frente a *1 - Especificidad* acumuladas de las n primeras predicciones. La especificidad establece la capacidad del clasificador de rechazar casos negativos y se calcula como:

$$Especificidad = \frac{TN}{TN + FP},$$

donde, *TN* corresponde a los casos no predichos que no interaccionan (para un determinado umbral de predicción del método). Por otro lado, la sensibilidad considera la capacidad del clasificador para detectar como positivos los casos realmente positivos y se calcula como:

$$Sensibilidad = \frac{TP}{TP + FN},$$

donde *FN* corresponde a las interacciones funcionales no predichas.

La representación de los valores de *Sensibilidad* frente a *1 – Especificidad* para diferentes niveles de confianza de un clasificador da lugar a una curva, cuyo área es una buena medida de la calidad del clasificador. Dicho área está acotada entre 0 y 1, correspondiendo un valor de 0,5 al esperado por azar. En este trabajo se realizó una curva *ROC* para las primeras 50.000 predicciones de *MT* y *CM-10*, respecto al conjunto *LT_COMPLEX*.

6.1.4.C Extracción de ejemplos ilustrativos

Los casos a estudiar en detalle fueron seleccionados manualmente a partir de una inspección visual de las predicciones más fiables. Sin embargo, con el fin de proporcionar una información completa de las relaciones co-evolutivas para los sistemas moleculares estudiados, se recuperaron todas las señales de co-evolución ($\rho' \geq 0,6$ y $p\ valor \leq 10^{-6}$) para cada proteína de cada sistema estudiado. De esta forma se puede observar en qué medida las proteínas de un sistema molecular tienden a co-evolucionar según *CM* con proteínas de otros sistemas.

6.2. Materiales y métodos asociados a los análisis realizados con *ContextMirror Global*

6.2.1. Base de datos de genomas secuenciados y selección de genomas

6.2.1.A NCBI Microbial Genome Resources (MGR)

MGR (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/) es un repositorio de genomas microbianos que pretende anotar los genomas provenientes de los experimentos de secuenciación. *MGR* contiene 4.338 genomas procariotas completos, de los cuales 4.132 son bacterianos. Este repositorio mantiene una selección de genomas anotados con sus herramientas para la predicción de genes en genomas procariotas. Así mismo, contiene información sobre los genomas anotados que facilita su uso. Todos los proteomas utilizados en los análisis realizados con *CMG* provienen de este recurso.

6.2.1.B. Selección de genomas

Los estudios realizados con *CMG* corresponden a una serie de 23 análisis paralelos centrados en 23 especies diferentes seleccionadas por *MICROME* (http://www.microme.eu/). A diferencia del análisis centrado en *E. coli* estos análisis fueron planteados para detectar la señal co-evolutiva específica de cada uno de los grupos taxonómicos a los que pertenecen estas especies. Así se empezó por clasificar los genomas bacterianos según su grupo taxonómico (ver Tabla 3).

Tabla 3. Genomas utilizados en los grupos taxonómicos de las 23 especies analizadas.

Grupo taxonómico	Número de proteomas
Actinobacteria	137
Alphaproteobacteria	147
Betaproteobacteria	99
Gammaproteobacteria	292
Deltaproteobacteria	42
Epsilonproteobacteria	39
Firmicutes	302
Bacteroidetes/Chlorobi group	66

El principal objetivo de esta clasificación es simplificar los cálculos y limitar las búsquedas por homología, ya que los criterios que realmente determinan la proximidad evolutiva de las especies incluidas se establecen sobre estos grupos.

En los análisis de *CMG*, se utilizaron las asignaciones a grupos taxonómicos para refinar la selección de especies. Así, en primer lugar se realizaron las búsquedas por homología para la especie de referencia frente a las especies de su grupo taxonómico usando el programa *BLAST* (Altschul *et al.*, 1997). De forma similar a lo descrito para los análisis con *ContextMirror*, se empleó la estrategia *BBH* entre las proteínas de cada una de las especies de referencia y las especies del grupo taxonómico al que pertenece dicha especie. Igualmente, se utilizaron los umbrales de *e-value* (10^{-5}) y de solapamiento en secuencia (70%) para seleccionar aquellos ortólogos más fiables.

Utilizando estas asignaciones de ortólogos y sus alineamientos recuperados por *BLAST*, se calculó para cada par de especies la identidad media entre sus ortólogos. Esta identidad media entre ortólogos es una medida aproximada de la distancia evolutiva entre las especies y se calcula como:

$$\langle \%IDort \rangle_{a,b} = \frac{\sum_{i=1}^{n} \left(\frac{I_{i,a,b}}{L_{i,a,b}} \right)}{n_{a,b}},$$

donde *a* y *b* corresponden a las especies comparadas, *n* es el número de asignaciones de ortología entre ambas especies, *I* es el número de posiciones idénticas entre un par de ortólogos y *L* corresponde a la longitud del alineamiento entre ambas secuencias.

A continuación se establecieron una serie de criterios para seleccionar los conjuntos de especies de cada análisis. Se eliminaron aquellas especies *a* que cumplieran alguna de las siguientes condiciones:

i) $n_{a,ref} < 1.000$;

ii) $\langle \%IDort \rangle_{a,ref} < 20\%$;

iii) $\langle \%IDort \rangle_{a,ref} > 80\%$;

iv) $\langle \%IDort \rangle_{a,b} > 80\%$ y $n_{a,ref} < n_{b,ref}, \forall a \neq ref, b \neq ref$,

donde *ref* designa a la especie de referencia. Estas condiciones están diseñadas para recuperar un conjunto de especies con unas divergencias entre ellas razonablemente homogéneas, lo que mejora la calidad de las señales recuperadas por nuestras metodologías (Herman *et al.*, 2011). La aplicación de estos criterios permitió recuperar conjuntos de especies adecuadas para los análisis con *CM* y *CMG* en las 23 especies estudiadas (ver Tabla 1).

6.2.2. Obtención de las matrices de distancias filogenéticas para las proteínas de las 23 especies de referencia

El protocolo para obtener las matrices distancias para las especies de referencia es el mismo que el explicado en la sección 6.1.2. El único cambio fue la introducción de *MAFFT* (Katoh & Toh, 2008) un programa de construcción de alineamientos múltiples que sustituye a *MUSCLE*. Este cambio viene motivado por una mejor calidad en los alineamientos múltiples obtenidos por *MAFFT* (Thompson *et al.*, 2011).

6.2.3. Construcción de los conjuntos de evaluación

En la comparación con *CM*, los conjuntos *LT_PPI*, *LT_COMPLEX* y *KEGG_PWY* son los mismos definidos en la sección 6.1.3.

*6.2.3.A. **STRING** (Szklarczyk et al., 2015)*

STRING (http://string-db.org/) es una base de datos que almacena información sobre asociaciones funcionales entre proteínas. *STRING* incluye interacciones físicas obtenidas de otras bases de datos, rutas metabólicas, relaciones de co-expresión, así como información extraída automáticamente de la literatura, asociaciones predichas por homología, perfiles filogenéticos y fusión génica. La inclusión de métodos predictivos permite a *STRING* proporcionar información, aunque menos fiable, para más de 1.100 organismos. De hecho, *STRING* establece relaciones funcionales para las 23 especies de referencia utilizadas proporcionando una puntuación que evalúa la fiabilidad de cada asociación funcional. Esta puntuación integra la información de todas estas fuentes en una inferencia de presencia de un par de proteínas en la misma ruta metabólica de *KEGG* (Mering *et al.*, 2005). Además, *STRING* también incorpora un protocolo bastante elaborado para la transferencia de anotaciones entre especies diferentes (Szklarczyk *et al.*, 2015). Por lo tanto, las asociaciones de *STRING* se pueden considerar como una extensión del conjunto *KEGG_PWY*, pero que también contiene asociaciones basadas en la acumulación de otras evidencias. En estos análisis se consideraron aquellas asociaciones funcionales con una puntuación mayor o igual a 900, que corresponden al conjunto de casos de la mayor confianza.

6.2.3.B. Construcción de los conjuntos de negativos

Los conjuntos de pares de proteínas sin asociación funcional en *STRING* para cada especie de referencia se obtuvieron siguiendo la misma estrategia que la descrita en la sección 6.1.3.D.

6.2.4 Evaluación de las predicciones obtenidas por *ContextMirror Global*

6.2.4.A. Evaluación de la precisión predictiva de ContextMirror Global

En los análisis con *ContextMirror Global* la evaluación comparativa de los distintos métodos en las distintas especies se realizó de la misma forma que se detalla en la

sección 6.1.4.A, con la única diferencia que en esta ocasión se utilizó para ello el paquete *ROCR* (Sing *et al.*, 2005) (versión 1.0-7) de *R* (R Core Team, 2013).

6.2.4.B. Análisis de los términos de ontología génica (GO) (Ashburner et al., 2000)

Para detectar los términos *GO* enriquecidos en la co-evolución detectada por *CMG* en cada una de las 23 especies estudiadas, se seleccionaron las proteínas implicadas en las primeras 500 predicciones confirmadas por *STRING*. La intención de esta decisión es establecer conjuntos de pares de proteínas comparables entre distintas especies que nos permitan estudiar su co-evolución con la menor influencia posible de las diferencias en las condiciones de cada análisis.

A continuación se recuperaron de *UniProt-GOA* (Barrell *et al.*, 2009) las anotaciones correspondientes a cada una de las especies de referencia. En este punto es importante tener en cuenta, que una buena parte de las anotaciones recuperadas han sido inferidas electrónicamente. Aunque esta es una situación inevitable en el contexto de un análisis que incluye tantas especies relativamente poco estudiadas, este hecho debe tenerse presente y ser cautos con las interpretaciones muy detalladas.

A continuación, se realizaron los análisis de enriquecimiento de términos *GO* basados en estas anotaciones utilizando *GO-TermFinder* (Boyle *et al.*, 2004). *GO-TermFinder* es un programa que establece *p valores* obtenidos a partir de la distribución hipergeométrica y los corrige posteriormente por test múltiple, siguiendo la aproximación de Bonferroni (Dunn, 1961). *GO-TermFinder* también estima las tasas de falsos descubrimientos (*FDR*). De esta forma se obtuvieron los términos *GO* sobre-representados (p *valor* < 0.01 y FDR $< 1\%$) en las proteínas con interacciones detectadas por *CMG* en cada una de las 23 especies.

Finalmente, se procedió a combinar los términos *GO* procedentes de diferentes especies para observar qué anotaciones aparecían más comúnmente representadas entre los pares que co-evolucionan en ellas. Ésta no es una tarea trivial, ya que las anotaciones pueden representar diferentes grados de especificidad en función de la calidad de las anotaciones o de las proteínas concretas implicadas en cada caso. Por lo tanto, se decidió utilizar *REVIGO* (Supek *et al.*, 2011), un recurso que permite obtener una visión global de análisis de *GO* complejos. Para ello se empleó el número de veces que cada anotación aparece sobre-representada en las 23 especies como medida de su importancia en estos análisis. *REVIGO* resume un conjunto de anotaciones en una selección de las más representadas y específicas Esto permite recuperar un número razonable de anotaciones que reflejan la estructura de los resultados. De entre las representaciones visuales que proporciona *REVIGO* se eligió la denominada *TreeMap* (Supek *et al.*, 2011). *TreeMap* agrupa las anotaciones obtenidas por *REVIGO* en otras anotaciones más generales, generando una jerarquía de dos niveles de especificidad de *GO*. La representación propiamente consiste en un gráfico rectangular cuya área se reparte entre las anotaciones en función de la importancia de la anotación (en este caso es proporcional al número de especies donde se detecta). Así, de los dos niveles de anotación recuperados por *REVIGO*, el más general se reparte el área con diferentes colores, mientras que las anotaciones del nivel más específico se reparten el área de la anotación general que las agrupa.

6.2.4.C. Extracción de ejemplos ilustrativos

En este caso, siguiendo la filosofía de la comparación entre especies, se seleccionaron los ejemplos entre aquellos sobre-representados en los análisis de *GO*. Con la intención de explorar cómo la co-evolución detectada se manifiesta en estos procesos en

diferentes especies, se recuperaron las asignaciones de sus proteínas a las rutas de *KEGG* más directamente relacionadas con las anotaciones de *GO* más interesantes. *KEGG* contiene información para todas las especies estudiadas en este trabajo excepto para *Pedobacter saltans*. Por lo tanto, para las restantes 22 especies, se pudieron obtener de *KEGG* todas las proteínas implicadas en las rutas seleccionadas (*Oxidative phosphorylation*, *Flagellar assembly* y *Transporters*). Para cada una de esta rutas se obtuvieron por separado las interacciones funcionales recuperadas por *CMG* entre las 500 predicciones de cada especie utilizadas en las sección anterior.

7. Bibliografía

Abby, S. S. & Rocha, E. P. C. (2012). The non-flagellar type III secretion system evolved from the bacterial flagellum and diversified into host-cell adapted systems. *PLoS Genetics*, 8(9), p e1002983.

Albert, R., Jeong, H. & Barabasi, A. (2000). Error and attack tolerance of complex networks. *Nature*, 406(6794), pp 378–382.

Althoff, D. M., Segraves, K. A. & Johnson, M. T. J. (2013). Testing for coevolutionary diversification: linking pattern with process. *Trends in ecology & evolution (Personal edition)*, pp 1–8 Elsevier Ltd.

Altschuh, D., Lesk, A. M., Bloomer, A. C. & Klug, A. (1987). Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of Molecular Biology*, 193(4), pp 693–707.

Altschuh, D., Vernet, T., Berti, P., Moras, D. & Nagai, K. (1988). Coordinated amino acid changes in homologous protein families. *Protein engineering*, 2(3), pp 193–199.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), pp 3389–3402.

Andrés-León, E., Ezkurdia, I., García, B., Valencia, A. & Juan, D. (2009). EcID. A database for the inference of functional interactions in E. coli. *Nucleic Acids Research*, 37(Database issue), pp D629–35.

Arifuzzaman, M. (2006). Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Research*, 16(5), pp 686–691.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1), pp 25–29.

Bader, G. D. & Hogue, C. W. V. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotechnology*, 20(10), pp 991–997.

Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T. & Hogue, C. W. (2001). BIND--The Biomolecular Interaction Network Database. *Nucleic Acids Research*, 29(1), pp 242–245.

Baldi, P., Brunak, S., Frasconi, P., Soda, G. & Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics (Oxford, England)*, 15(11), pp 937–946.

Bange, G., Kümmerer, N., Engel, C., Bozkurt, G., Wild, K. & Sinning, I. (2010). FlhA provides the adaptor for coordinated delivery of late flagella building blocks to the type III secretion system. *Proceedings of the National Academy of Sciences*, 107(25), pp 11295–11300.

Barrell, D., Dimmer, E., Huntley, R. P., Binns, D., O'Donovan, C. & Apweiler, R. (2009). The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Research*, 37(Database issue), pp D396–403.

Bascompte, J., Jordano, P., Melián, C. J. & Olesen, J. M. (2003). The nested assembly of plant-animal mutualistic networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16), pp 9383–9387.

Bastolla, U., Fortuna, M. A., Pascual-García, A., Ferrera, A., Luque, B. & Bascompte, J. (2009). The architecture of mutualistic networks minimizes competition and increases

biodiversity. *Nature*, 458(7241), pp 1018–1020.

Bauer, B., Mirey, G., Vetter, I. R., García-Ranea, J. A., Valencia, A., Wittinghofer, A., Camonis, J. H. & Cool, R. H. (1999). Effector recognition by the small GTP-binding proteins Ras and Ral. *The Journal of biological chemistry*, 274(25), pp 17763–17770.

Becker, A.-K., Zeppenfeld, T., Staab, A., Seitz, S., Boos, W., Morita, T., Aiba, H., Mahr, K., Titgemeyer, F. & Jahreis, K. (2006). YeeI, a novel protein involved in modulation of the activity of the glucose-phosphotransferase system in Escherichia coli K-12. *The Journal of Bacteriology*, 188(15), pp 5439–5449.

Berger, A. L., Pietra, V. & Pietra, S. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, 22, pp 39–71.

Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M. & Sherlock, G. (2004). GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18), pp 3710–3715.

Brandt, U. (2006). Energy converting NADH:quinone oxidoreductase (complex I). *Annual Review of Biochemistry*, 75, pp 69–92.

Brown, B. M., Wang, Z., Brown, K. R., Cricco, J. A. & Hegg, E. L. (2004). Heme O synthase and heme A synthase from Bacillus subtilis and Rhodobacter sphaeroides interact in Escherichia coli. *Biochemistry*, 43(42), pp 13541–13548.

Brown, J. K. M. & Tellier, A. (2011). Plant-parasite coevolution: bridging the gap between genetics and ecology. *Annual review of phytopathology*, 49, pp 345–367.

Browne, W. J., North, A. C., Phillips, D. C., Brew, K., Vanaman, T. C. & Hill, R. L. (1969). A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. *Journal of Molecular Biology*, 42(1), pp 65–86.

Butland, G., Peregrín-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J. & Emili, A. (2005). Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature*, 433(7025), pp 531–537.

Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J. & Huang, E. S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein science : a publication of the Protein Society*, 13(1), pp 190–202.

Carmona, D., Fitzpatrick, C. R. & Johnson, M. T. J. (2015). 50 years of coevolution and beyond: Integrating coevolution from molecules to species. *Molecular Ecology*.

Carro, A., Tress, M., de Juan, D., Pazos, F., Lopez-Romero, P., del Sol, A., Valencia, A. & Rojas, A. M. (2006). TreeDet: a web server to explore sequence space. *Nucleic Acids Research*, 34(Web Server), pp W110–W115.

Casari, G., Sander, C. & Valencia, A. (1995). A method to predict functional residues in proteins. *Nature structural biology*, 2(2), pp 171–178.

Chen, S., Beeby, M., Murphy, G. E., Leadbetter, J. R., Hendrixson, D. R., Briegel, A., Li, Z., Shi, J., Tocheva, E. I., Müller, A., Dobro, M. J. & Jensen, G. J. (2011). Structural diversity of bacterial flagellar motors. *The EMBO Journal*, 30(14), pp 2972–2981.

Chuang, H.-Y., Hofree, M. & Ideker, T. (2010). A decade of systems biology. *Annual review of cell and developmental biology*, 26, pp 721–744.

Cocco, S., Monasson, R. & Weigt, M. (2013). From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction. Supporting Information. *PLoS Computational Biology*, (5), pp 1–21.

Contini, A. & Tiana, G. (2015). A many-body term improves the accuracy of effective potentials based on protein coevolutionary data. *The Journal of chemical physics*, 143(2), p 025103.

Cott, H. B. (1940). *Adaptive Colouration in Mammals*. Methuen.

Cox, J. & Mann, M. (2011). Quantitative, high-resolution proteomics for data-driven systems biology. *Annual Review of Biochemistry*, 80, pp 273–299.

Cramér, H. (1999). *Mathematical methods of statistics*. 19. ed Princeton University Press.

Currie, C. R., Mueller, U. G. & Malloch, D. (1999). The agricultural pathology of ant fungus gardens. *Proceedings of the National Academy of Sciences of the United States of America*, 96(14), pp 7998–8002.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection*. London: Murray.

Darwin, C. (1862). *On the various contrivances by which British and foreign orchids are fertilised by insects: and on the good effects of intercrossing*. London: John Murray.

Degnan, J. H. & Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2(5), p e68.

Degnan, J. H. & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24(6), pp 332–340.

del Sol Mesa, A., Pazos, F. & Valencia, A. (2003). Automatic Methods for Predicting Functionally Important Residues. *Journal of Molecular Biology*, 326(4), pp 1289–1302.

DeLuna, A., Vetsigian, K., Shoresh, N., Hegreness, M., Colón-González, M., Chao, S. & Kishony, R. (2008). Exposing the fitness contribution of duplicated genes. *Nature Genetics*, 40(5), pp 676–681.

Deutch, C. E., Spahija, I. & Wagner, C. E. (2014). Susceptibility of Escherichia coli to the toxic L-proline analogue L-selenaproline is dependent on two L-cystine transport systems. *Journal of applied microbiology*, 117(5), pp 1487–1499.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293), pp 52–64.

Dunn, S. D., Wahl, L. M. & Gloor, G. B. (2008). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3), pp 333–340.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5, p 113.

Edgar, R. S., Green, E. W., Zhao, Y., van Ooijen, G., Olmedo, M., Qin, X., Xu, Y., Pan, M., Valekunja, U. K., Feeney, K. A., Maywood, E. S., Hastings, M. H., Baliga, N. S., Merrow, M., Millar, A. J., Johnson, C. H., Kyriacou, C. P., O'Neill, J. S. & Reddy, A. B. (2012). Peroxiredoxins are conserved markers of circadian rhythms. *Nature*, 485(7399), pp 459–464.

Ehrlich, P. & Raven, P. (1964). Butterflies and plants: a study in coevolution. *Evolution*, 18(4), pp 586–608.

Eichler, W. (1948). Some rules in ectoparasitism
. *Annals and Magazine of Natural History (Series 12)*, 1, pp 588–598.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 87(1), p 012707.

Fahrenholz, H. (1913). Ectoparasiten unde abstammungslehre. *Zoologische Anzeiger (Leipzig)*, (41), pp 371–374.

Fares, M. A. & McNally, D. (2006). CAPS: coevolution analysis using protein sequences. *Bioinformatics*, 22(22), pp 2821–2822.

Fariselli, P., Olmea, O., Valencia, A. & Casadio, R. (2001). Prediction of contact maps with neural networks and correlated mutations. *Protein engineering*, 14(11), pp 835–843.

Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Research*, 40(Database issue), pp D136–43.

Feinauer, C., Skwark, M. J., Pagnani, A. & Aurell, E. (2014). Improving contact prediction along three dimensions. *PLoS Computational Biology*, 10(10), p e1003847.

Filizola, M., Olmea, O. & Weinstein, H. (2002). Prediction of heterodimerization interfaces of G-protein coupled receptors with a new subtractive correlated mutation method. *Protein engineering*, 15(11), pp 881–885.

Fischer, D. (2000). Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pacific Symposium on Biocomputing 2012*, pp 119–130.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic zoology*, 19(2), pp 99–113.

Flicek, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Gräf, S., Haider, S., Hammond, M., Holland, R., Howe, K. L., Howe, K., Johnson, N., Jenkinson, A., Kähäri, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A. J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A. & Searle, S. (2008). Ensembl 2008. *Nucleic Acids Research*, 36(Database issue), pp D707–14.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Overduin, B., Pritchard, B., Riat, H. S., Rios, D., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sobral, D., Spudich, G., Tang, Y. A., Trevanion, S., Vandrovcova, J., Vilella, A. J., White, S., Wilder, S. P., Zadissa, A., Zamora, J., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Vogel, J. & Searle, S. M. J. (2011). Ensembl 2011. *Nucleic Acids Research*, 39(Database issue), pp D800–6.

Fodor, A. A. & Aldrich, R. W. (2004). Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins: Structure, Function, and Bioinformatics*, 56(2), pp 211–221.

Foote, R. (2007). Mathematics and Complex Systems. *Science*, 318(5849), pp 410–412.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L. & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4), pp 1531–1545.

Fox, L. R. (1981). Defense and dynamics in plant-herbivore systems. *American Zoologist*, (21), pp 853–864.

Friedrich, T. & Scheide, D. (2000). The respiratory complex I of bacteria, archaea and eukarya and its module common with membrane-bound multisubunit hydrogenases. *FEBS Letters*, 479(1-2), pp 1–5.

Ganmor, E., Segev, R. & Schneidman, E. (2011). Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proceedings of the National Academy of Sciences*, 108(23), pp 9679–9684.

García-Jiménez, B., Juan, D., Ezkurdia, I., Andrés-León, E. & Valencia, A. (2010). Inference of functional relations in predicted protein networks with a machine learning approach.(García-Jiménez, B., Juan, D., Ezkurdia, I., Andrés-León, E., & Valencia, A., Eds) *PloS one*, 5(4), p e9969.

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. & Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868), pp 141–147.

Gertz, J., Elfond, G., Shustrova, A., Weisinger, M., Pellegrini, M., Cokus, S. & Rothschild, B. (2003). Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics*, 19(16), pp 2039–2045.

Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D. & Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. *Journal of Molecular Biology*, 299(2), pp 283–293.

González-Pedrajo, B., Minamino, T., Kihara, M. & Namba, K. (2006). Interactions between C ring proteins and export apparatus components: a possible mechanism for facilitating type III protein export. *Molecular microbiology*, 60(4), pp 984–998.

Gophna, U., Ron, E. Z. & Graur, D. (2003). Bacterial type III secretion systems are ancient and evolved by multiple horizontal-transfer events. *Gene*, 312, pp 151–163.

Göbel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 18(4), pp 309–317.

Guttman, L. (1938). A note on the derivation of formulae for multiple and partial correlation. *Annals of Mathematical Statistics*, 9(4), pp 305–308.

Hafner, M. S. & Nadler, S. A. (1988). Phylogenetic trees support the coevolution of parasites and their hosts. *Nature*, 332(6161), pp 258–259.

Hajirasouliha, I., Schönhuth, A., Juan, D., Valencia, A. & Sahinalp, S. C. (2012). Mirroring co-evolving trees in the light of their topologies. *Bioinformatics*, 28(9), pp 1202–1208.

Harrell, E. J. (2015). Hmisc: Harrell Miscellaneous. Available from: http://CRAN.R-project.org/package=Hmisc.

Hayat, S., Sander, C., Marks, D. S. & Elofsson, A. (2015). All-atom 3D structure prediction of transmembrane β-barrel proteins from sequences. *Proceedings of the National Academy of Sciences*, 112(17), pp 5413–5418.

Hembry, D. H., Yoder, J. B. & Goodman, K. R. (2014). Coevolution and the Diversification of Life. *The American naturalist*, 184(4), pp 425–438.

Herman, D., Ochoa, D., Juan, D., Lopez, D., Valencia, A. & Pazos, F. (2011). Selection of organisms for the co-evolution-based study of protein interactions. *BMC Bioinformatics*, 12(1), p 363.

Hernanz-Falcón, P., Rodríguez-Frade, J. M., Serrano, A., Juan, D., del Sol, A., Soriano, S. F., Roncal, F., Gómez, L., Valencia, A., Martínez-A, C. & Mellado, M. (2004). Identification of amino acid residues crucial for chemokine receptor dimerization. *Nature Immunology*, 5(2), pp 216–223.

Hirano, T., Minamino, T. & Macnab, R. M. (2001). The role in flagellar rod assembly of the N-terminal domain of Salmonella FlgJ, a flagellum-specific muramidase. *Journal of Molecular Biology*, 312(2), pp 359–369.

Hogenesch, J. B. & Ueda, H. R. (2011). Understanding systems-level properties: timely stories from the study of clocks. *Nature Reviews Genetics*, 12(6), pp 407–416.

Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C. & Marks, D. S. (2012). Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*,

149(7), pp 1607–1621.

Hopf, T. A., Morinaga, S., Ihara, S., Touhara, K., Marks, D. S. & Benton, R. (2015). Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. *Nature Communications*, 6 SP -, p 6077 Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.

Hopf, T. A., Schärfe, C. P. I., Rodrigues, J. P. G. L. M., Green, A. G., Kohlbacher, O., Sander, C., Bonvin, A. M. J. J. & Marks, D. S. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes.(Kuriyan, J., Ed) *eLife Sciences* [online], 3, pp 1–45. Available from: http://elifesciences.org/content/3/e03430.abstract.

Innan, H. & Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2), pp 97–108 Nature Publishing Group.

Izarzugaza, J. M. G., del Pozo, A., Vazquez, M. & Valencia, A. (2012). Prioritization of pathogenic mutations in the protein kinase superfamily. *BMC genomics*, 13 Suppl 4, p S3.

Izarzugaza, J. M. G., Hopcroft, L. E. M., Baresic, A., Orengo, C. A., Martin, A. C. R. & Valencia, A. (2011). Characterization of pathogenic germline mutations in human protein kinases. *BMC Bioinformatics*, 12 Suppl 4, p S1.

Izarzugaza, J. M. G., Juan, D., Pons, C., Ranea, J. A. G., Valencia, A. & Pazos, F. (2006). TSEMA: interactive prediction of protein pairings between interacting families. *Nucleic Acids Research*, 34(Web Server), pp W315–W319.

Izarzugaza, J. M., Juan, D., Pons, C., Pazos, F. & Valencia, A. (2008). Enhancing the prediction of protein pairings between interacting families using orthology information. *BMC Bioinformatics*, 9(1), p 35.

Jacob, F., Perrin, D., Sanchez, C. & Monod, J. (1960). [Operon: a group of genes with the expression coordinated by an operator]. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 250, pp 1727–1729.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. II. *Physical Review Series II*, 108(2), pp 171–190.

Jaynes, E. T. (1957b). Information theory and statistical mechanics. *Physical Review Series II*, 106(4), pp 620–630.

Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology*, 287(4), pp 797–815.

Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2), pp 184–190.

Jothi, R., Kann, M. G. & Przytycka, T. M. (2005). Predicting protein-protein interaction by searching evolutionary tree automorphism space. *Bioinformatics*, 21 Suppl 1, pp i241–50.

Joyce, A. R. & Palsson, B. Ø. (2006). The model organism as a system: integrating "omics" data sets. *Nature Reviews Molecular Cell Biology*, 7(3), pp 198–210.

Juan, D., Graña, O., Pazos, F., Fariselli, P., Casadio, R. & Valencia, A. (2003). A neural network approach to evaluate fold recognition results. *Proteins: Structure, Function, and Bioinformatics*, 50(4), pp 600–608.

Juan, D., Mellado, M., Rodríguez-Frade, J. M., Hernanz-Falcón, P., Serrano, A., del Sol, A., Valencia, A., Martínez-A, C. & Rojas, A. M. (2005). A framework for computational and experimental methods: identifying dimerization residues in CCR chemokine receptors. *Bioinformatics*, 21 Suppl 2, pp ii13–8.

Juan, D., Pazos, F. & Valencia, A. (2008a). Co-evolution and co-adaptation in protein networks. *FEBS Letters*, 582(8), pp 1225–1230.

Juan, D., Pazos, F. & Valencia, A. (2008b). High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proceedings of the National Academy of Sciences of the United States of America*, 105(3), pp 934–939.

Juan, D., Pazos, F. & Valencia, A. (2013). Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4), pp 249–261.

Juan, D., Perner, J., Carrillo-de Santa Pau, E., Marsili, S., Ochoa, D., Chung, H.-R., Vingron, M., Rico, D. & Valencia, A. (2015). Epigenomic co-localization and co-evolution reveal a key role for 5hmC as a communication hub in the chromatin network of ESCs. *Submitted.*

Jukes, T. H. (1963). Some recent advances in studies of the transcription of the genetic message. *Advances in biological and medical physics*, 9, pp 1–41.

Kalinina, O. V., Mironov, A. A., Gelfand, M. S. & Rakhmaninova, A. B. (2004). Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein science : a publication of the Protein Society*, 13(2), pp 443–456.

Kandel, E. R., Dudai, Y. & Mayford, M. R. (2014). The molecular and systems biology of memory. *Cell*, 157(1), pp 163–186.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(suppl 1), pp D277–D280.

Katoh, K. & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, 9(4), pp 286–298.

Katsonis, P. & Lichtarge, O. (2014). A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. *Genome Research*, 24(12), pp 2050–2058.

Kellogg, V. L. (1896). New Mallophaga, 1. With special reference to a collection from maritime birds of the Bay of Monterey, California. *Proceedings of the California Academy of Science*, (6), pp 31–168.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R. & Hermjakob, H. (2007). IntAct--open source resource for molecular interaction data. *Nucleic Acids Research*, 35(Database), pp D561–D565.

Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E. & Apweiler, R. (2004). The International Protein Index: an integrated database for proteomics experiments. *PROTEOMICS*, 4(7), pp 1985–1988.

Kersey, P. J., Lawson, D., Birney, E., Derwent, P. S., Haimel, M., Herrero, J., Keenan, S., Kerhornou, A., Koscielny, G., Kähäri, A., Kinsella, R. J., Kulesha, E., Maheswari, U., Megy, K., Nuhn, M., Proctor, G., Staines, D., Valentin, F., Vilella, A. J. & Yates, A. (2010). Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Research*, 38(Database issue), pp D563–9.

Kersey, P., Bower, L., Morris, L., Horne, A., Petryszak, R., Kanz, C., Kanapin, A., Das, U., Michoud, K., Phan, I., Gattiker, A., Kulikova, T., Faruque, N., Duggan, K., Mclaren, P., Reimholz, B., Duret, L., Penel, S., Reuter, I. & Apweiler, R. (2005). Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Research*, 33(Database issue), pp D297–302.

Keseler, I. M. (2004). EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Research*, 33(Database issue), pp D334–D337.

Kihara, M., Minamino, T., Yamaguchi, S. & Macnab, R. M. (2001). Intergenic suppression

between the flagellar MS ring protein FliF of Salmonella and FlhA, a membrane component of its export apparatus. *The Journal of Bacteriology*, 183(5), pp 1655–1662.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2), pp 111–120.

Kinoshita, M., Hara, N., Imada, K., Namba, K. & Minamino, T. (2013). Interactions of bacterial flagellar chaperone-substrate complexes with FlhA contribute to co-ordinating assembly of the flagellar filament. *Molecular microbiology*, 90(6), pp 1249–1261.

Kojima, S. & Blair, D. F. (2004). Solubilization and purification of the MotA/MotB complex of Escherichia coli. *Biochemistry*, 43(1), pp 26–34.

Kolde, R. (2015). pheatmap: Pretty Heatmaps. Available from: http://CRAN.R-project.org/package=pheatmap.

Kondrashov, F., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. (2002). Selection in the evolution of gene duplications. *Genome Biology*, 3(2), p RESEARCH0008.

Korber, B. T., Farber, R. M., Wolpert, D. H. & Lapedes, A. S. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 90(15), pp 7176–7180.

Köster, U., Sohl-Dickstein, J., Gray, C. M. & Olshausen, B. A. (2014). Modeling higher-order correlations within cortical microcolumns. *PLoS Computational Biology*, 10(7), p e1003684.

Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Hoad, G., Kanz, C., Lee, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Nardone, F., Pastor, M. P. G., Plaister, S., Sobhany, S., Stoehr, P., Vaughan, R., Wu, D., Zhu, W. & Apweiler, R. (2007). EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Research*, 35(Database issue), pp D16–20.

Lapedes, A. S., Giraud, B. G. & Jarzynski, C. (2002). Using Sequence Alignments to Predict Protein Structure and Stability With High Accuracy. pp 1–29. Available from: http://arxiv.org/abs/1207.2484.

Lapedes, A. S., Giraud, B. G., Liu, L. & Stormo, G. D. (1999). Correlated Mutations in Models of Protein Sequences: Phylogenetic and Structural Effects. *Lecture Notes-Monograph Series*, 33 IS -, pp 236–256 Institute of Mathematical Statistics.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21), pp 2947–2948. Lawrence, J. G. (2002). Shared strategies in gene organization among prokaryotes and eukaryotes. *Cell*, 110(4), pp 407–413.

Ledoit, O. & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10, pp 603–621.

Li, X., Romero, P., Rani, M., Dunker, A. & Obradovic, Z. (1999). Predicting Protein Disorder for N-, C-, and Internal Regions. *Genome informatics. Workshop on Genome Informatics*, 10, pp 30–40.

Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology*, 257(2), pp 342–358.

Lichtarge, O., Yao, H., Kristensen, D. M., Madabushi, S. & Mihalek, I. (2003). Accurate and scalable identification of functional sites by evolutionary tracing. *Journal of structural and functional genomics*, 4(2-3), pp 159–166.

Liu, R. & Ochman, H. (2007). Stepwise formation of the bacterial flagellar system. *Proceedings of the National Academy of Sciences of the United States of America*, 104(17), pp 7116–7121.

Lynch, M. & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494), pp 1151–1155.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R. & Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PloS one*, 6(12), p e28766.

Marston, M. F., Pierciey, F. J., Shepard, A., Gearin, G., Qi, J., Yandava, C., Schuster, S. C., Henn, M. R. & Martiny, J. B. H. (2012). Rapid diversification of coevolving marine Synechococcus and a virus. *Proceedings of the National Academy of Sciences of the United States of America*, 109(12), pp 4544–4549.

Martin, L. C., Gloor, G. B., Dunn, S. D. & Wahl, L. M. (2005). Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 21(22), pp 4116–4124.

Mathiesen, C. & Hägerhäll, C. (2003). The "antiporter module" of respiratory chain complex I includes the MrpC/NuoK subunit -- a revision of the modular evolution scheme. *FEBS Letters*, 549(1-3), pp 7–13.

McMurry, J. L., Van Arnam, J. S., Kihara, M. & Macnab, R. M. (2004). Analysis of the cytoplasmic domains of Salmonella FlhA and interactions with components of the flagellar export machinery. *The Journal of Bacteriology*, 186(22), pp 7586–7592.

McPartland, J. M., Norris, R. W. & Kilpatrick, C. W. (2007). Coevolution between cannabinoid receptors and endocannabinoid ligands. *Gene*, 397(1-2), pp 126–135.

Mering, von, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A. & Bork, P. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(Database issue), pp D433–7.

Mering, von, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887), pp 399–403.

Meyer, J. R., Dobias, D. T., Weitz, J. S., Barrick, J. E., Quick, R. T. & Lenski, R. E. (2012). Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science*, 335(6067), pp 428–432.

Mézard, M., Parisi, G. & Virasoro, M. (1986). *Spin Glass Theory and Beyond*. World Scientific. ISBN 978-981-279-937-1.

Mihalek, I., Res, I. & Lichtarge, O. (2004). A family of evolution-entropy hybrid methods for ranking protein residues by importance. *Journal of Molecular Biology*, 336(5), pp 1265–1282.

Minamino, T. & Macnab, R. M. (2000). Interactions among components of the Salmonella flagellar export apparatus and its substrates. *Molecular microbiology*, 35(5), pp 1052–1064.

Minamino, T., Kinoshita, M., Hara, N., Takeuchi, S., Hida, A., Koya, S., Glenwright, H., Imada, K., Aldridge, P. D. & Namba, K. (2012). Interaction of a bacterial flagellar chaperone FlgN with FlhA is required for efficient export of its cognate substrates. *Molecular microbiology*, 83(4), pp 775–788.

Montoya, J. M., Pimm, S. L. & Solé, R. V. (2006). Ecological networks and their fragility. *Nature*, 442(7100), pp 259–264.

Moparthi, V. K. & Hägerhäll, C. (2011). The evolution of respiratory chain complex I from

a smaller last common ancestor consisting of 11 protein subunits. *Journal of Molecular Evolution*, 72(5-6), pp 484–497.

Morange, M. & Cobb, M. (2000). *A History of Molecular Biology*. Harvard University Press. ISBN 9780674001695.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. E. N., Hwa, T. & Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49), pp E1293–301.

Morillas, M., Gómez-Puertas, P., Bentebibel, A., Sellés, E., Casals, N., Valencia, A., Hegardt, F. G., Asins, G. & Serra, D. (2003). Identification of conserved amino acid residues in rat liver carnitine palmitoyltransferase I critical for malonyl-CoA inhibition. Mutation of methionine 593 abolishes malonyl-CoA inhibition. *The Journal of biological chemistry*, 278(11), pp 9058–9063.

Nambu, T., Minamino, T., Macnab, R. M. & Kutsukake, K. (1999). Peptidoglycan-hydrolyzing activity of the FlgJ protein, essential for flagellar rod formation in Salmonella typhimurium. *The Journal of Bacteriology*, 181(5), pp 1555–1561.

Neher, E. (1994). How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences of the United States of America*, 91(1), pp 98–102.

Ochoa, D., García-Gutiérrez, P., Juan, D., Valencia, A. & Pazos, F. (2013). Incorporating information on predicted solvent accessibility to the co-evolution-based study of protein interactions. *Molecular BioSystems*, 9(1), pp 70–76 The Royal Society of Chemistry.

Ochoa, D., Juan, D., Valencia, A. & Pazos, F. (2015). Detection of significant protein coevolution. *Bioinformatics*, 31(13), pp 2166–2173.

Ohno, S. (1970). *Evolution by gene duplication*. Springer-Verlag.

Oparina, N. J., Kalinina, O. V., Gelfand, M. S. & Kisselev, L. L. (2005). Common and specific amino acid residues in the prokaryotic polypeptide release factors RF1 and RF2: possible functional implications. *Nucleic Acids Research*, 33(16), pp 5226–5234.

Opgen-Rhein, R. & Strimmer, K. (2007). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statistical applications in genetics and molecular biology*, 6(1), pp –.

Osman, A. A., Monroe, M. M., Ortega Alves, M. V., Patel, A. A., Katsonis, P., Fitzgerald, A. L., Neskey, D. M., Frederick, M. J., Woo, S. H., Caulin, C., Hsu, T.-K., McDonald, T. O., Kimmel, M., Meyn, R. E., Lichtarge, O. & Myers, J. N. (2015). Wee-1 kinase inhibition overcomes cisplatin resistance associated with high-risk TP53 mutations in head and neck cancer through mitotic arrest followed by senescence. *Molecular cancer therapeutics*, 14(2), pp 608–619.

Ovchinnikov, S., Kamisetty, H. & Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife*, 3, p e02030.

Pallen, M. J. & Matzke, N. J. (2006). From The Origin of Species to the origin of bacterial flagella. *Nature reviews. Microbiology*, 4(10), pp 784–790.

Pallen, M. J., Beatson, S. A. & Bailey, C. M. (2005). Bioinformatics, genomics and evolution of non-flagellar type-III secretion systems: a Darwinian perspective. *FEMS microbiology reviews*, 29(2), pp 201–229.

Parkinson, J. S., Parker, S. R., Talbert, P. B. & Houts, S. E. (1983). Interactions between chemotaxis genes and flagellar genes in Escherichia coli. *Journal of Bacteriology*, 155(1), pp 265–274.

Parrish, J. K. & Edelstein-Keshet, L. (1999). Complexity, pattern, and evolutionary trade-

offs in animal aggregation. *Science*, 284(5411), pp 99–101.

Paterson, S., Vogwill, T., Buckling, A., Benmayor, R., Spiers, A. J., Thomson, N. R., Quail, M., Smith, F., Walker, D., Libberton, B., Fenton, A., Hall, N. & Brockhurst, M. A. (2010). Antagonistic coevolution accelerates molecular evolution. *Nature*, 464(7286), pp 275–278.

Pauling, L. & Zuckerkandl, E. (1963). Chemical paleogenetics: molecular "restoration studies" of extinct forms of life. *Acta Chemica Scandinavica*, 17, pp 9–16.

Pazos, F. & Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein engineering*, 14(9), pp 609–614.

Pazos, F. & Valencia, A. (2002). In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins: Structure, Function, and Bioinformatics*, 47(2), pp 219–227.

Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *Journal of Molecular Biology*, 271(4), pp 511–523.

Pazos, F., Ranea, J. A. G., Juan, D. & Sternberg, M. J. E. (2005). Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *Journal of Molecular Biology*, 352(4), pp 1002–1015.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. In: *Proceedings of Proceedings of the Royal Society of London*, 1895. pp 240–242.

Peterson, S. M., Pack, T. F., Wilkins, A. D., Urs, N. M., Urban, D. J., Bass, C. E., Lichtarge, O. & Caron, M. G. (2015). Elucidation of G-protein and β-arrestin functional selectivity at the dopamine D2 receptor. *Proceedings of the National Academy of Sciences*, 112(22), pp 7097–7102.

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., Gabaldón, T., Rattei, T., Creevey, C., Kuhn, M., Jensen, L. J., Mering, von, C. & Bork, P. (2014). eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Research*, 42(Database issue), pp D231–9.

Pritchard, L. & Birch, P. (2011). A systems biology perspective on plant-microbe interactions: biochemical and structural targets of pathogen effectors. *Plant science : an international journal of experimental plant biology*, 180(4), pp 584–603.

Pujol, A., Mosca, R., Farrés, J. & Aloy, P. (2010). Unveiling the role of network and systems biology in drug discovery. *Trends in Pharmacological Sciences*, 31(3), pp 115–123.

Punta, M. & Rost, B. (2005). PROFcon: novel prediction of long-range contacts. *Bioinformatics (Oxford, England)*, 21(13), pp 2960–2968.

Ramani, A. K. & Marcotte, E. M. (2003). Exploiting the co-evolution of interacting proteins to discover interaction specificity. *Journal of Molecular Biology*, 327(1), pp 273–284.

Rausell, A., Juan, D., Pazos, F. & Valencia, A. (2010). Protein interactions and ligand binding: From protein subfamilies to functional specificity. *Proceedings of the National Academy of Sciences* [online], 107(5), pp 1995–2000 National Academy of Sciences. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2808218\&tool=pmcentrez\&rendertype=abstract.

Reddy, T. B. K., Thomas, A. D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., Mallajosyula, J., Pagani, I., Lobos, E. A. & Kyrpides, N. C. (2015). The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Research*, 43(Database issue), pp D1099–106.

Reva, B., Antipin, Y. & Sander, C. (2007). Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biology*, 8(11), p R232.

Rodriguez, G. J., Yao, R., Lichtarge, O. & Wensel, T. G. (2010). Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. *Proceedings of the National Academy of Sciences of the United States of America*, 107(17), pp 7787–7792.

Rojas, A. M., Fuentes, G., Rausell, A. & Valencia, A. (2012). The Ras protein superfamily: evolutionary tree and role of conserved amino acids. *The Journal of Cell Biology*, 196(2), pp 189–201.

Rosenberg, N. A. & Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, 3(5), pp 380–390.

Rost, B. & Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proceedings of the National Academy of Sciences*, 90(16), pp 7558–7562.

Rost, B., Casadio, R., Fariselli, P. & Sander, C. (1995). Transmembrane helices predicted at 95% accuracy. *Protein science : a publication of the Protein Society*, 4(3), pp 521–533.

Saiki, K., Mogi, T., Hori, H., Tsubaki, M. & Anraku, Y. (1993a). Identification of the functional domains in heme O synthase. Site-directed mutagenesis studies on the cyoE gene of the cytochrome bo operon in Escherichia coli. *The Journal of biological chemistry*, 268(36), pp 26927–26934.

Saiki, K., Mogi, T., Ogura, K. & Anraku, Y. (1993b). In vitro heme O synthesis by the cyoE gene product from Escherichia coli. *The Journal of biological chemistry*, 268(35), pp 26041–26044.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4), pp 406–425.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. & Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, 32(Database issue), pp D449–51.

Sato, T., Yamanishi, Y., Kanehisa, M. & Toh, H. (2005). The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, 21(17), pp 3482–3489.

Schäfer, J. & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4, pp –.

Schäfer, J., Opgen-Rhein, R., Zuber, V., Ahdesmäki, M., Duarte Silva, P. & Strimmer, K. (2015). corpcor: Efficient Estimation of Covariance and (Partial) Correlation. Available from: http://CRAN.R-project.org/package=corpcor.

Schlesinger, D. H. & Goldstein, G. (1975). Molecular conservation of 74 amino acid sequence of ubiquitin between cattle and man. *Nature*, 255(5507), pp 423–424.

Schlicker, A., Domingues, F. S., Rahnenführer, J. & Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7, p 302.

Schut, G. J., Boyd, E. S., Peters, J. W. & Adams, M. W. W. (2013). The modular respiratory complexes involved in hydrogen and sulfur metabolism by heterotrophic hyperthermophilic archaea and their evolutionary implications. *FEMS microbiology reviews*, 37(2), pp 182–203.

Scott-Phillips, T. C. (2008). Defining biological communication. *Journal of evolutionary biology*, 21(2), pp 387–395.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N.,

Schwikowski, B. & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), pp 2498–2504.

Shenoy, S. K., Drake, M. T., Nelson, C. D., Houtz, D. A., Xiao, K., Madabushi, S., Reiter, E., Premont, R. T., Lichtarge, O. & Lefkowitz, R. J. (2006). beta-arrestin-dependent, G protein-independent ERK1/2 activation by the beta2 adrenergic receptor. *The Journal of biological chemistry*, 281(2), pp 1261–1273.

Shindyalov, I. N., Kolchanov, N. A. & Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein engineering*, 7(3), pp 349–358.

Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics (Oxford, England)*, 21(20), pp 3940–3941.

Smith, J. M. & Harper, D. (2003). *Animal Signals*. Oxford University Press. ISBN 9780198526858.

Sokal, R. R. & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, 11, pp 33–40.

Stein, R. R., Marks, D. S. & Sander, C. (2015). Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. *PLoS Computational Biology*, 11(7), p e1004182.

Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. *Journal of Molecular Evolution*, 49(2), pp 169–181.

Student (1908). The Probable Error of a Mean. *Biometrika*, 6(1), pp 1–25.

Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one*, 6(7), p e21800.

Sułkowska, J. I., Morcos, F., Weigt, M., Hwa, T. & Onuchic, J. N. (2012). Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences*, 109(26), pp 10340–10345.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J. & Mering, von, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(Database issue), pp D447–52.

Tansley, A. G. (1935). The Use and Abuse of Vegetational Concepts and Terms. *Ecology ...*, 16(3), pp 284–307 Ecological Society of America.

Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N. P., Hayes, W. S., Borodovsky, M., Rudd, K. E. & Koonin, E. V. (1996). Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with Escherichia coli. *Current biology : CB*, 6(3), pp 279–291.

Taylor, W. R. & Hatrick, K. (1994). Compensating changes in protein multiple sequence alignments. *Protein engineering*, 7(3), pp 341–348.

R Core Team. (2013). R: A Language and Environment for Statistical Computing. Available from: http://www.R-project.org/.

Thébault, E. & Fontaine, C. (2010). Stability of ecological communities and the architecture of mutualistic and trophic networks. *Science*, 329(5993), pp 853–856.

Thompson, J. D., Linard, B., Lecompte, O. & Poch, O. (2011). A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PloS one*, 6(3), p e18093.

Thompson, J. N. (1994). *The coevolutionary process*. University Of Chicago Press. ISBN 9780226797595.

Thompson, J. N. (2005). *The Geographic Mosaic of Coevolution*. University of Chicago Press. ISBN 0226797627.

Thompson, J. N. (2013). *Relentless Evolution*. University of Chicago Press. ISBN 022601889X.

Thrall, P. H., Hochberg, M. E., Burdon, J. J. & Bever, J. D. (2007). Coevolution of symbiotic mutualists and parasites in a community context. *Trends in ecology & evolution (Personal edition)*, 22(3), pp 120–126.

Tillier, E. R. M. & Lui, T. W. H. (2003). Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics*, 19(6), pp 750–755.

Tiwary, B., Tiwary, B. K. & Li, W.-H. (2009). Parallel evolution between aromatase and androgen receptor in the animal kingdom. *Molecular Biology and Evolution*, 26(1), pp 123–129.

Tkacik, G., Marre, O., Mora, T., Amodei, D., Berry, M. J., II & Bialek, W. (2013). The simplest maximum entropy model for collective behavior in a neural network. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03), p P03011.

Tress, M., Juan, D., Graña, O., Gómez, M. J., Gómez-Puertas, P., González, J. M., López, G. & Valencia, A. (2005). Scoring docking models with evolutionary information.(Janin, J., Ed) *Proteins: Structure, Function, and Bioinformatics*, 60(2), pp 275–280.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. & Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770), pp 623–627.

UniProt Consortium (2007). The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 35(Database issue), pp D193–7.

Van Valen, L. (1973). A new evolutionary law. *Evolutionary theory*, 1(1), pp 1–30.

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(1), pp 67–72.

Werner, H. M. J., Mills, G. B. & Ram, P. T. (2014). Cancer Systems Biology: a peek into the future of patient care? *Nature reviews. Clinical oncology*, 11(3), pp 167–176.

Whittaker, J. (2009). *Graphical Models in Applied Multivariate Statistics*. Wiley. ISBN 9780470743669.

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S.-M. & Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1), pp 303–305.

Zaman, L., Meyer, J. R., Devangam, S., Bryson, D. M., Lenski, R. E. & Ofria, C. (2014). Coevolution drives the emergence of complex traits and promotes evolvability.(Keller, L., Ed) *PLoS Biology* [online], 12(12), p e1002023. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=25514332&retmode=ref&cmd=prlinks.

Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M. & Cesareni, G. (2002). MINT: a Molecular INTeraction database. *FEBS Letters*, 513(1), pp 135–140.

Anexo. Artículos publicados por el doctorando relacionados con la tesis

A.1. Artículos en el área de la co-evolución entre proteínas

A.2. Artículos en el área de la co-evolución entre residuos de aminoácidos

A.3. Artículos acerca de recursos web y bases de datos

A.4. Revisiones en el campo de la co-evolución molecular